Deep learning based skill assessment in transoesophageal echocardiography

Kamakshi Bansal

Supervised by

Dr. Danail Stoyanov, UCL

Dr. Evangelos Mazomenos, UCL

MSc Data Science

September 2017

Abstract

Transoesophageal Echocardiography (TEE) is a diagnostic test of the heart which requires echocardiographers to have high level of psychomotor skills: (a) maneuvering the TEE probe safely and fluently via the esophagus to the desired location in the heart; (b) getting appropriate images of the desired planes of the heart. Traditionally, these technical skills have been assessed manually. In this project, for the first time, deep learning based frameworks called as Convolutional Neural Networks (CNNs) have been used to build an artificially intelligent system to evaluate the heart images captured by of an echocardiographer during TEE.

Experiments were performed to acquire the ultrasound heart images from 10 suggested views of the heart using HeartWorks TEE simulator. Each of these images were evaluated manually with two different scores, criteria-based percentage score (0% to 100%) and a subjective general impression score (0 to 4). Two established CNN networks, AlexNet and VGG16, were modified and fine-tunned with the heart images to perform regression over each type of the two scores mentioned above and thereby predict the manual score. These networks were also modified to classify each heart image into ten defined planes of the heart. In general three models per network, two regression models and one classification model, were built. These models were experimented with the original heart images as well as its four different processed versions for detailed investigation. The training was done on GPU, for 44 different combinations of models and images.

The Root Mean Square Error (RMSE) for predicting criteria percentage score was in the range 16.80-19.10 % for AlexNet and 9.64-12.51% for VGG16 . The RMSE for predicting general impression score was in range 0.77-0.87 for AlexNet

and 0.40-0.56 for VGG16. The classification model for AlexNet and VGG16 gave similar accuracies i.e. 98.021%. and 98.332% respectively. The source code can be found at https://github.com/kamakshi22/Skill-assessment-TEE-DeepLearing.

Acknowledgements

Firstly, I would like to thank my academic supervisor, Dr. Danail Stoyanov, for giving me an opportunity to work on this project. Secondly, many thanks to my internal supervisor, Dr. Evangelos Mazomenos for supporting me throughout the project consistently. The guidance and ideas given by him and his excellent insight helped in shaping the final outcome of the project. Furthermore, I would like to thank the members of the Surgical Robotic Vision Team of UCL who always made me feel welcomed and as a part of the team. My thanks also go to everyone who supported me through thick and thin while working on this project.

Contents

1 Introduction **13**

1.1 Dissertation objectives 15

1.2 Structure of the dissertation 16

2 Background and related work **17**

2.1 Transesophageal Echocardiography (TEE) 17

2.1.1 TEE probe handling 19

2.1.2 Guidelines for performing TEE 21

2.2 TEE skills 22

2.2.1 Virtual reality surgical simulators 24

2.3 Machine learning 25

2.4 Neural networks and deep learning approaches 26

2.4.1 Neural networks 26

2.4.2 Convolutional neural networks (CNN) 29

2.4.3 Layers of CNN 31

2.4.4 Backpropagation 36

2.4.5 Deep learning architectures 37

2.4.6 Transfer learning and fine-tuning 37

2.4.7 Why CNNs are better? 38

2.4.8 Applications of CNN in medical imaging 39

3 Methods and experiments **43**

3.1 Data Preprocessing 44

3.1.1 Data acquisition from HeartWorks VR simulator 44
3.1.2 Manual data annotation 47
3.2 Video preprocessing 49
3.3 Image preprocessing 49
3.4 CNN architectures 53
3.4.1 AlexNet 55
3.4.2 VGG16 58
3.5 Dataset preparation 59
3.6 Experiments performed 59

4 Results and discussions **61**
4.1 Loss graphs obtained after training the CNNs 61
4.2 Results obtained from different processed versions of original images 62
4.3 Comparison of scoring techniques 64
4.4 Comparison of different networks 66
4.5 Results obtained from the shuffled dataset 66

5 Conclusions and future work **70**
5.1 Summary of the work 70
5.2 Conclusions 71
5.3 Future research directions 73

Bibliography **75**

List of Figures

2.1 Positioning of transducer in TTE and TEE. Sketch taken from [1]. . 17

2.2 Image showing how a transoesophageal echocardiography (TEE) is performed. Image adapted from [2]. 18

2.3 Image showing how a standard echocardiogram (TTE) is performed. Image adapted from[3]. 19

2.4 TEE probe manipulation. Image adapted from [4]. 20

2.5 Image depicting TEE probe insertion and image scanning. Image adapted from [5]. 20

2.6 The 20 recommended views for performing TEE by the ASE and ASC societies. Image taken from [6]. 22

2.7 Kinematics of state-of-the-art probe movement assessment. Taken from [7]. 24

2.8 A practitioner carrying out TEE via HeartWorks virtual reality simulator. Image adapted from [5]. 25

2.9 A perceptron. Image adapted from [8]. 27

2.10 A sketch of a neural network. Image taken from [8]. 27

2.11 The workflow required for training a neural network. Image adapted from [8]. 29

2.12 Filters after convolving over image generates respective feature maps. Image adapted from [9] 30

2.13 Systematic workflow of convolutional netural networks showing CNN layers. Image adapted from [10]. 31

2.14 (a) Represents the filter; (b) Represents an input image over which filter is convolved; (c) The resultant activation map. Image adapted from [11] 32

2.15 As the filter is sliding through the image, each of the filter weights gets multiplied with the corresponding pixel value of the image and sum of all these multiplications is stored in the top left corner of the generated feature map. Image adapted from [10]. 32

2.16 An example showing stride of 1. Image adapted from [12] 33

2.17 (a) Input image of size 32 x 32 x 3 ; (b) Block of size 26 x 26 x 10 obtained after convolving 10 filters of size 7 x 7 x 3 over an input image; (c) Block of size 11 x 11 x 6 obtained after convolving 6 filters of size 5 x 5 x 10. Image adapted from [11] 34

2.18 Image showing Zero Padding. Image adapted from [11] 34

2.19 The non linear activation function RELU, $f(x) = max(0,x)$. Image adapted from [13] 35

2.20 Image representing the two pooling techniques: average pooling and max pooling. Image adapted from [8] 36

2.21 Image showing CNN classifying the images with the presence and absence of skin cancer. Image adapted from [14]. 40

2.22 Image showing segmentation of Left Atrium, the yellow contour enclosing heart in a Cardiac Magnetic Resonance Imaging (MRI).[15] 40

2.23 Image showing an example of object detection. The MRI images on the left side are of the breast of two patients with confirmed cancer (done by biopsy). The red arrows points the area which is proved to be malignant. The images on the right hand side are the predictions done by CNN network recognizing the suspicious findings (red region) with respect to the images on the left. Image adapted from [16] 41

3.1 Image showing the architecture followed step by step to achieve the objective of the project. 44

3.2 A trainee experimenting with HeartWorks TEE simulator. Image adapted from [5]. 45

3.3 A participant performing TEE for the experiments done for the acquisition of data. Image adapted from [5]. 45

3.4 (a) Represents 3D rendering of the heart model; (b) Represents the corresponding simulated ultrasound view. Image adapted from [5]. . 46

3.5 A snippet from an excel sheet showing the marking done for plane number 1 for one of the participants. 47

3.6 A snippet of the CSV file generated. 48

3.7 A sample frame obtained after processing a video (heart plane number 1) [5]. 50

3.8 The image obtained after processing the frame shown in (original image frame). 50

3.9 Canny image obtained using canny edge detector algorithm [17]. . . 51

3.10 Enhanced image obtained after morphological processing with white hate filter and Otsu thresholding [17]. 51

3.11 Segmentation contour image obtained using active contours segmentation algorithm [17]. 52

3.12 Segmentation image after filling the above segmentation contour image [17]. 52

3.13 Overfitting prevention using dropout. Image adapted from [18] . . . 55

3.14 AlexNet Architecture used for performing classification and regression in the project. The last fully connected layer fc8 is modified according to the problem under consideration. Further, a softmax layer is added when performing classification to classify an image to any one of the ten defined planes of the heart and a dense layer with one neuron is added when performing regression using each of the manual scores. Original architecture adapted from [19]. 57

3.15 VGG16 Architecture used for performing classification and regression. The last fully connected layer fc8 is modified according to the problem under consideration. Further, a softmax layer is added when performing classification to classify an image to any one of the ten defined planes of the heart and a dense layer with one neuron is added when performing regression using each of the manual scores. Original architecture adapted from [20]. 58

4.1 The first and second image represents the tensorboard loss graphs for the regression performed on original images (using AlexNet) with criteria percentage score and general impression score as labels respectively. 61

4.2 The first and the second image represents the tensorboard loss graphs for the regression performed on original images (using VGG16) with criteria percentage score and general impression score as labels respectively. 62

4.3 Results from AlexNet showing that the performance of regression over criteria percentage score is better than regression over general impression score. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively. 64

4.4 Results from VGG16 showing that the performance of regression over criteria percentage score is better than regression over general impression score for all the frames except segmentation contour images. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively. 65

4.5 The first and the second image represents the tensorboard loss graphs for the regression performed on original images (using AlexNet) with criteria percentage score and general impression score as labels using the dataset shuffled over images respectively. . 67

4.6 The first and the second image represents the tensorboard loss graphs for the regression performed on original images (using VGG16) with criteria percentage score and general impression score as labels using the dataset shuffled over images respectively. . 67

4.7 Results from AlexNet using the dataset shuffled over images. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively. 68

4.8 Results from VGG16 using the dataset shuffled over images. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively. The performance of regression over criteria percentage score is better than regression over general impression score. 69

5.1 (a) Imaging plane; (b) 3D model; (c) 2D TEE image; (d) Acquisition protocol; (e) Structures imaged for Mid-Esophageal 4-Chamber (centered at tricuspid valve) - ME4C (TV) (plane number 1). Image adapted from [4] 72

5.2 (a) Imaging plane; (b) 3D model; (c) 2D TEE image; (d) Acquisition protocol; (e) Structures imaged for Mid-Esophageal Aortic Valve Short-Axis - ME AV SAX (plane number 3). Image adapted from [4] 72

List of Tables

3.1 Sequence of the 10 ultrasound image planes used in the study 46

4.1 The results obtained from AlexNet. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression. 63

4.2 The results obtained from VGG16. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression. 63

4.3 The results obtained from AlexNet using using the dataset shuffled over images. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression. 68

4.4 The results obtained from VGG16 using using the dataset shuffled over images. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression. 68

Chapter 1

Introduction

Transoesophageal echocardiography (TEE) is an established test for diagnosis of complications in the heart. TEE test is performed by inserting the TEE probe inside the esophagus, proximally to the heart for its scanning. This test requires an echocardiographer to have an exceptionally high level of technical skills for maneuvering the TEE probe, through the esophagus, fluently and safely to the desired location and taking the ultrasound images of the heart. To attain proficiency in these technical skills, proper training programme for an echocardiographers with performance feedback is required. Traditionally, the assessment of these skills is done manually by expert supervisors which is time consuming and laborious. Training procedure can be improved significantly if the automation of such skill assessments can be done which apart from saving time will also increase the assessment accuracy. Automating the assessment will also minimize the need of an expert supervisor as well as allow the trainee to repeat TEE, on a dummy torso, unlimited number of times at any time according to their availability and comfort. Though studies related to the evaluation of surgical skills on the basis of the motion analysis of the tip of the probe during TEE have been done [7], but there has been little research related to an automatic evaluation of the ultrasound heart images obtained by an echocardiographer during TEE. In this project, to build a system which could automatically evaluate the ultrasound heart images obtained from a TEE and give them a score which is near to the manual scores given by the experts, Convolutional Neural Networks (CNNs) have been used.

Recent advances of CNN in the field of computer vision, has prompted a surge of interest to apply similar set of techniques to medical images. Particularly, fine-tunning of existing deep learning architectures also known as transfer learning is very popular. Training of a CNN from scratch is a time consuming and tedious process as it requires large amount of annotated data which, in general is difficult to obtain in the context of medical imaging as expert annotation is expensive. In such situations, transfer learning technique is applied where existing pre-trained CNNs are fine-tuned to make it compatible with medical images being used. CNNs have been successfully used in the field of medical science for the diagnosis of various diseases like lung cancer, breast cancer, skin cancer etc. where they have even out-performed the human accuracy in detection of cancer [21]. In 2016, Menegola et al. successfully detected Melanoma (skin cancer) after training the dataset with 1000 images of skin lesions using transfer learning [22]. Hence, motivated by the popularity and impressive performance of CNNs in the medical field, a deep learning approach using them (CNNs) has been implemented in this project. This has been done for the first time and there has been no publication till date which investigates the automation of skill assessment in TEE using deep learning.

For implementing any deep learning network, preparation of the dataset is the foremost step. To acquire the data for this project, experiments were performed involving 38 participants (23 novices and 15 experts) where each participant was required to obtain the 10 suggested views of the heart using HeartWorks TEE simulator. Hence, there were ten types of the heart images obtained by each participant. Each type of the heart image had some criteria to be satisfied and it was manually evaluated by the expert anesthetists for the two different scores. First type of the score called as criteria percentage score, was evaluated from 0 to 100% and was based on the number of criteria to be full-filled by each participant while taking the heart image. The second type of the score called as general impression score, was evaluated on a scale of 0-4 purely on the basis of the experience and understanding of the experts.

After the preparation of dataset, the CNN models were built. To built the CNN

model which could predict each of the two types of the aforementioned manual scores, two established deep learning architectures AlexNet and VGG16 were modified to perform regression over each of the score. In other words, for both the networks one CNN model was built to do regression over criteria percentage score and another was built to perform regression over general impression score. Regression with CNNs was done to get a continuous score for a given image. In simple terms it was done as instead of classification of the pixels of an input image, prediction of a single value close to the manual score was required. Two classification models using each of the AlexNet and VGG16 were also built which could classify the heart image into the ten different defined planes of the heart. Hence, a total of 6 CNN models, 3 CNNs each for both AlexNet and VGG16 were constructed out of which two models were built to perform regression and one model was built to perform classification for each network.

The training of all the six CNN models built was first done with the original images of the heart obtained from the HeartWorks TEE simulator. For detailed investigation, the original images were further processed using four different kinds of image processing techniques to remove noise and specular highlights in the original image. These four kinds of images were mainly enhanced, canny, segmentation and segmentation contour. The training was done on GPU, graphical processing units which makes the computations faster, for 44 different combinations of models and images.

1.1 Dissertation objectives

The primary objective of this dissertation is to modify and implement the current state-of-the-art deep learning approaches to automate the assessment of imaging skills of echocardiographers carrying out the TEE. Or in other words, to build an automatic system which could replicate the manual scoring done by the experts while evaluating the imaging skills of the individuals performing TEE.

The additional objectives of this thesis are:

- To implement two well known deep learning architectures AlexNet and VGG16 and compare their performance.
- To investigate the performance of deep learning architectures on original images of the heart obtained from the HeartWorks TEE simulator as well as its four different types of processed versions.
- To investigate the potential of a CNN in predicting the manual scores similar to those given by the expert evaluators.

1.2 Structure of the dissertation

The structure of this thesis is as follows: Chapter Two explains Transoesophageal echocardiography (TEE) and the skills required for performing it efficiently. It further discusses the details and the functioning of convolutional neural networks (CNNs) and transfer learning. Chapter Three describes the methods implemented in achieving the aforementioned objectives. It also explains the experiments performed with the deep learning architectures, AlexNet and VGG16. Chapter Four presents the results achieved from the investigations done, compares the architectures implemented and discusses the reasons in detail. Finally, the future directions for the research work and the achievements and the work performed in this project have been concluded in Chapter Five.

Chapter 2

Background and related work

2.1 Transesophageal Echocardiography (TEE)

Transoesophageal echocardiography (TEE or TOE) is an important diagnostic procedure which provides ultrasound scans of the heart and the great vessels [7][4]. It is an excellent way to examine the proper functioning of the heart and finding anomalies (if any) in its structure. TEE scans the heart using an ultrasound transducer attached to the tip of a flexible endoscope (probe), navigated through the esophagus proximally to the heart (Fig. 2.1 (b)). There are many clinical applications of TEE such as evaluation of various congenital heart diseases (CHD) in both children and adults, detection of complications of endocarditis, aortic dissection, other aortic pathologies and evaluation of valvular disorders [23][24].

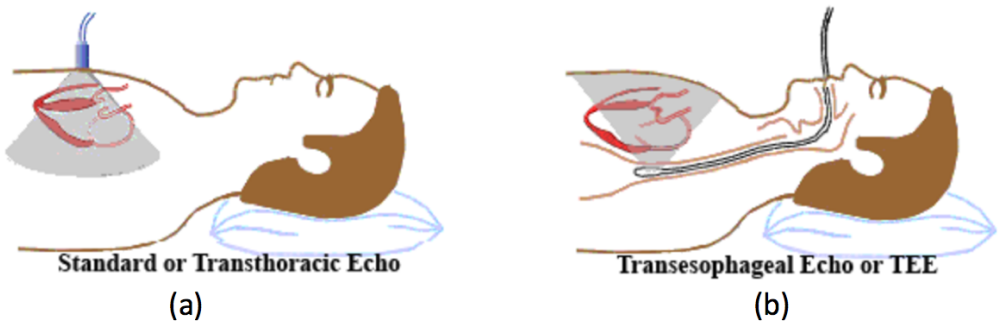


Figure 2.1: Positioning of transducer in TTE and TEE. Sketch taken from [1].

TEE is not limited to just these clinical applications, but also useful for the pa-

Figure 2.2: Image showing how a transoesophageal echocardiography (TEE) is performed. Image adapted from [2].

tients suspected with cardiac trauma, tumors inside the heart and in certain clinical situations of patients in which TTE (transthoracic echocardiogram) assessment is inadequate [25] [1]. (TTE is a standard echocardiogram done with the help of an ultrasound probe in a *transthoracic* way, meaning that the ultrasound waves travel through the front of the chest or thorax (Fig. 2.1 (a)) [1]. An ultrasound imaging jelly is applied to the chest area or to the ultrasound probe directly and the patient is asked to lie down and the probe is moved over the patient's chest (Fig. 2.3). This test usually lasts about 15-60 minutes and has no side effects [3].)

TEE provides more accurate and detailed images of the heart than TTE. It is because in TEE the ultrasound waves do not have to travel through the chest walls like in TTE and also heart lies immediately in front of the esophagus, which enables TEE probe to obtain a clearer and more detailed image (Fig. 2.2). Overall, TEE allows extraction of more accurate information of the four chambers and the valves of the heart, particularly the back structures such as the left atrium, which is usually not clearly visible in a TTE. Consequently a more accurate cardiovascular diagnosis and monitoring can be done through TEE [1].

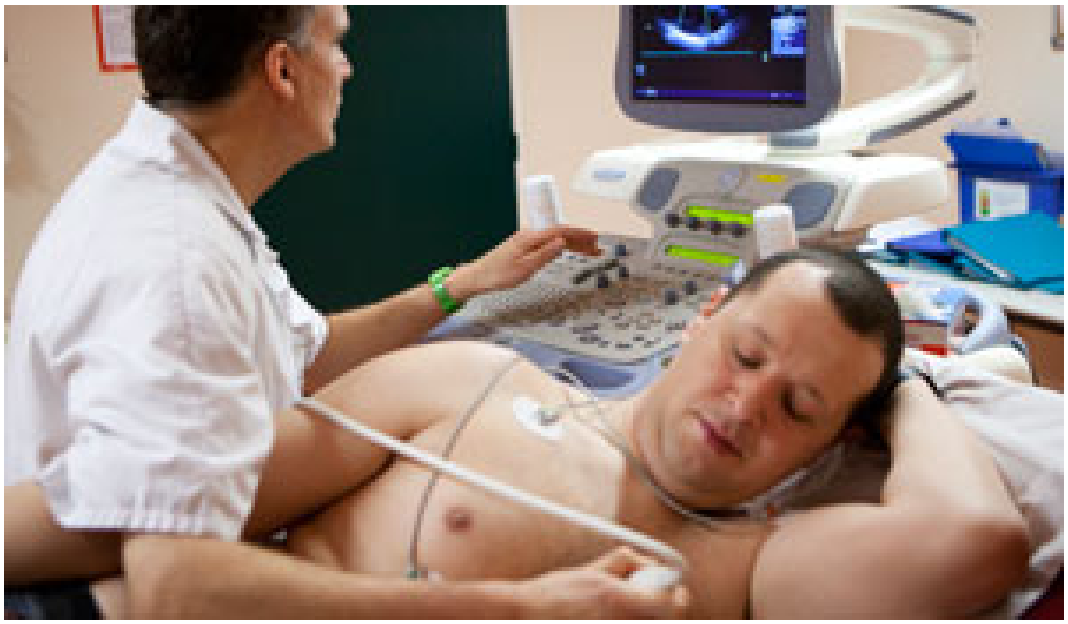


Figure 2.3: Image showing how a standard echocardiogram (TTE) is performed. Image adapted from[3].

A TEE procedure is performed after giving moderate sedation to the patient. However, patients requiring a TEE as part of a surgical procedure are fully anesthetized and intubated in general. Though the actual procedure only lasts for about 10-30 min, the overall procedure including all preparation and observations last for roughly 2-3 hours [1]. In general, TEE is a minimally invasive procedure, however, it involves slight risks (mentioned in section 2.1.1) and may cause discomfort to the patient [26] [27]. Therefore, it must be performed in those clinical circumstances where the TTE is inadequate for diagnosis.

2.1.1 TEE probe handling

The TEE probe, inserted in the esophagus to examine the heart, can be maneuvered in different directions to obtain the different views of the heart (as shown in Fig. 2.6). For example, it can be moved up and down the esophagus or its tip can be deflected or its scan plane may be rotated (as shown in Fig. 2.4). It emits ultrasonic waves which maps the approximated densities of all objects within its triangular shaped planar scanning region to the TEE screen (Fig. 2.5 (a)). The scanning region is then displayed as a fan shaped gray scale image on the monitor screen as

can be seen in Fig. 2.5 (b) [4].

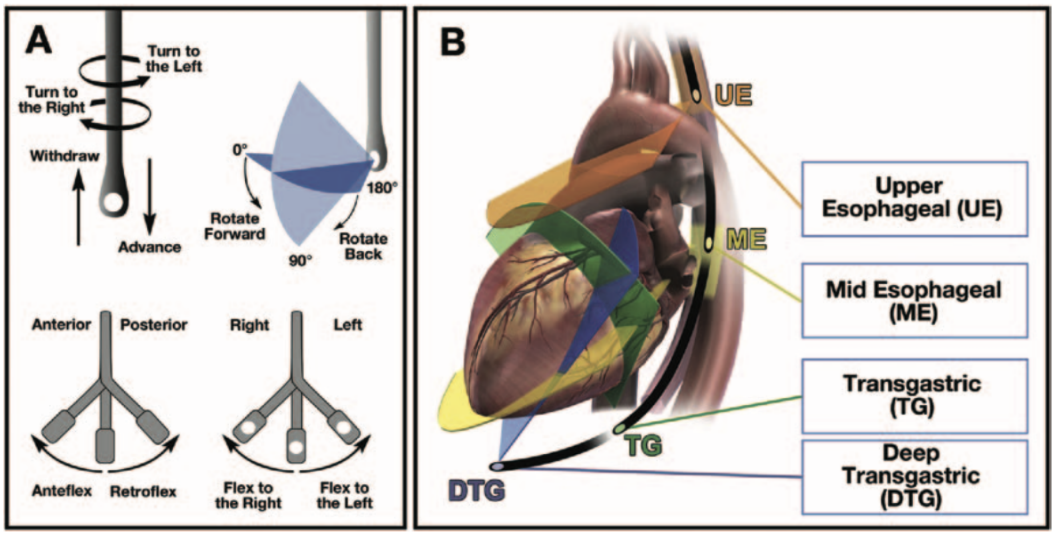
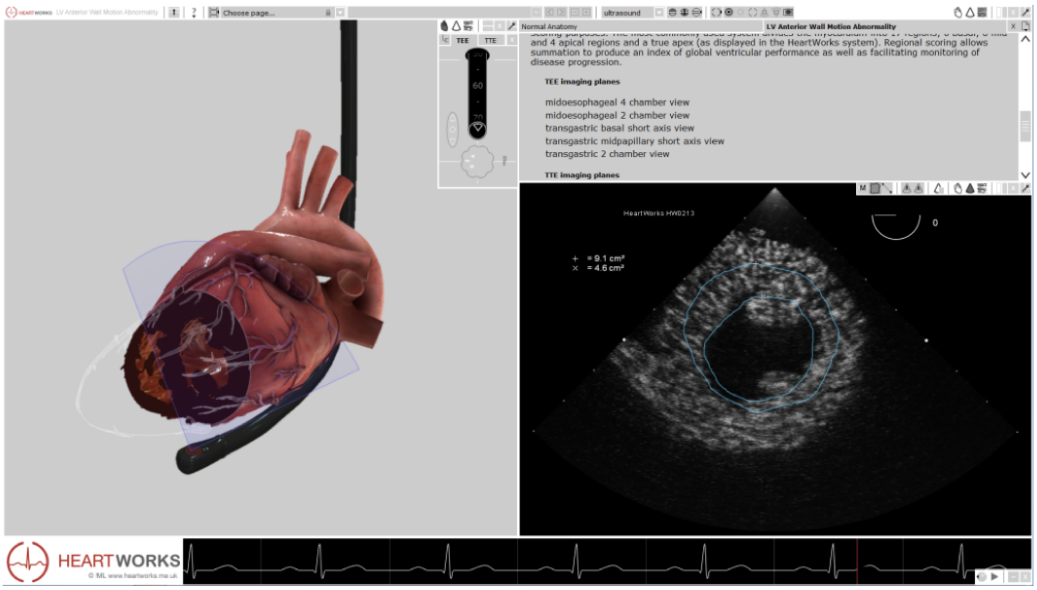


Figure 2.4: TEE probe manipulation. Image adapted from [4].



(a) (b)

Figure 2.5: Image depicting TEE probe insertion and image scanning. Image adapted from [5].

The TEE test must be performed by a practitioner who is an expert in handling the TEE probe. Although TEE is a safe procedure it might lead to some complications if not performed by a skilled echocardiographer [26] [27]. There could be

injuries in the gastrointestinal tract such as dental trauma, tonsillar bleeding etc. It may also lead to respiratory or cardiovascular complications and morbid infections such as endocarditis as described by [28]. In [28] the authors highlight a variety of heath related issues with the improper insertion of the probe.

The second problem which may arise is the incorrect usage of a TEE probe like scanning at incorrect locations or with wrong angulations. This will not give the required views of the heart as discussed in the next section [6]. Proper training with assessment and feedback is the best way to solve these two technical issues which is discussed in Sec. 2.2 [7][29]. There are guidelines published routinely by professional accreditation organizations for the proper training and interpretation of TEE as discussed in the upcoming section.

2.1.2 Guidelines for performing TEE

Performing a TEE test has become a standard practice in most cardiac procedures since it helps in diagnoses of heart anomalies with an excellent efficiency. The American Society of Echocardiography (ASE) and the Society of Cardiovascular Anesthesiologists (SCA) published guidelines for performing a comprehensive intra-operative multi-plane transesophageal echocardiographic examination [6][4]. These guidelines require an echocardiographer to record 20 specific views (Fig. 2.6) of the heart including all four chambers and valves of the heart, thoracic aorta and the pulmonary artery. In 2013, guidelines defining 28 views of the heart were introduced by the aforementioned societies [6]. Any echocardiographer usually has an individual approach to performing a TEE test and the order in which the images are taken during the test is not essential. The most important aspect of TEE is that it must be completed without leaving out any of the required views as mentioned in the guidelines [4] [6]. These guidelines also define certain cognitive and technical skills which a trainee must possess to perform the TEE. Few of those skills are discussed in the next section [4].

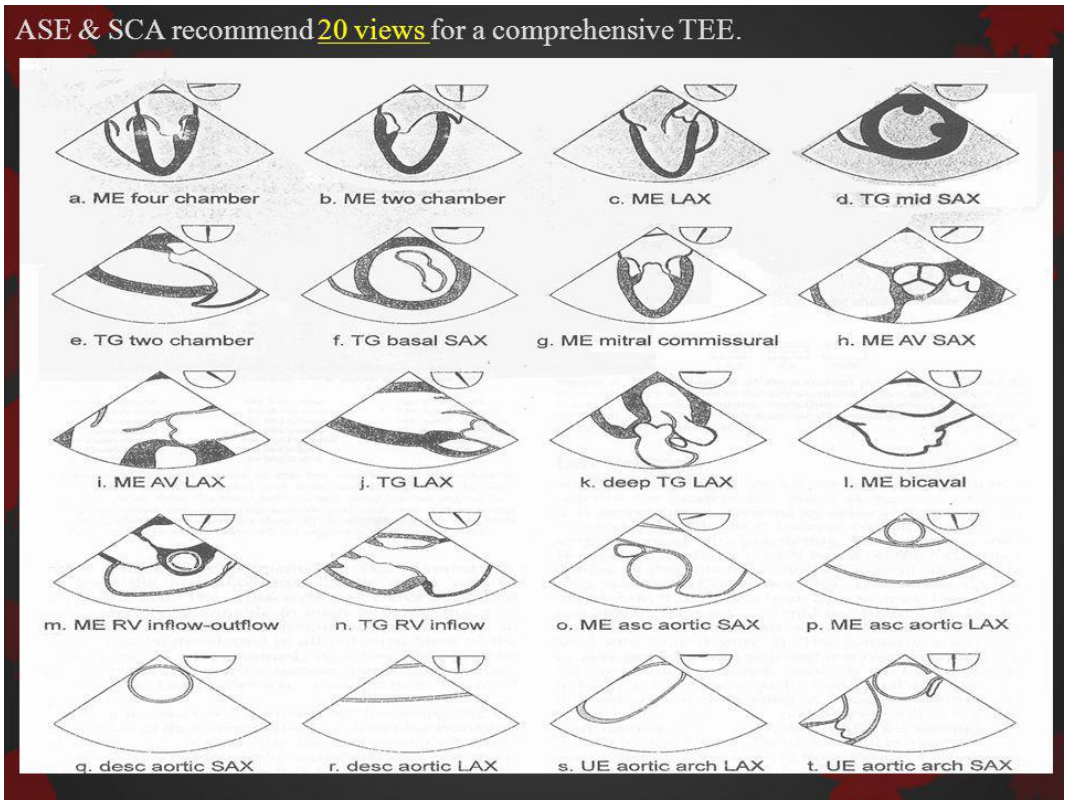


Figure 2.6: The 20 recommended views for performing TEE by the ASE and ASC societies. Image taken from [6].

2.2 TEE skills

The surgical skills vary from person to person depending on the time spent on training, teaching and their experiences so far [29]. These skills play an important role in surgical operations and have significant impact on the health care of the patients [30]. There are different types of factors which attribute to the skill level of a surgeon such as cognitive capabilities, manual dexterity, decision making and judgment [30][7]. To efficiently perform TEE, an electrocardiographer is required to undergo proper training and possess complex psychomotor skills and technical skills such as a)inserting TEE probe to the desired location safely and properly, b)manipulating and adjusting the probe correctly to get the accurate images of the heart and c)properly interpreting the results and communicating them to the other health-care providers in an eloquent and smooth manner [4]. Guidelines are published routinely by professionally accredited organizations as mentioned in

Sec. 2.1.2 for the basic training to attain the required cognitive and technical skills [6] [25].

The maneuvering of the probe to scan the correct views of the heart and inserting it safely to the desired location through the esophagus are the main technical skills required. The other skills mentioned above are the cognitive skills addressing the standard medical knowledge[7]. Cognitive skills can be taught in the classroom and evaluated in a written or an oral examination. These skills are best learned through an active participation in formal fellowship in cardiovascular medicine or its equivalent [25][29]. Evaluation and assessment of technical skills is a difficult process and needs to be standardized in order to guarantee a coherent outcome of surgical procedures throughout the world. The dexterity achieved for these TEE skills achieved by an echocardiographer depends heavily on the training and feedback system.

Traditionally, the training and assessment of a trainee is done under the supervision of a senior faculty surgeon who is also an experienced echocardiographer with significant expertise in TEE [25] [7]. This practice is time consuming, requires an expert supervision and has limited standardization [31]. A more standardized way to assess these skills needs to be formulated. Simulation based training using virtual reality simulators has been introduced as discussed in the Section 2.2.1, which allows novice endovascular interventionalists to achieve appropriate levels of competency regardless of their background [32]. A virtual reality surgical simulator replicates the surgical environment and allows trainees to practice TEE test unlimited number of times to hone their clinical and psychomotor skills without posing any risks to the patient [33][32]. These simulators help the trainees in surgical training outside the operating room environment by providing real time feedback [30].

Till now, the improvement of trainees during their training has only been assessed on the basis of how fluently they handled the probe and not on precision of the TEE views taken by them. To this end, Fig. 2.7 summarizes the precise parameters with which the handling of the probe was assessed. These results were summarized from the data collected using virtual reality simulator. Additionally,

Parameter	Novices	Experts	p-value (MW)
T_t - Total time (sec)	439.9	226.2	0.0007
pl - Depth path length (m)	2.838	2.509	0.482
v_d - Average depth velocity (m/s)	0.007	0.009	0.0004
a_d - Average depth acceleration (m/s^2)	0.240	0.304	0.009
j_d - Depth dimensionless jerk	15.938	2.353	0.0004
η_{sal} - Depth spectral arc length	−15.478	−6.684	0.028
tw - Average twist (deg)	18.832	11.906	0.0004
fl - Average flexion (deg)	6.748	5.619	1

Figure 2.7: Kinematics of state-of-the-art probe movement assessment. Taken from [7].

the proficiency of the skills of the novices and expert TEE echocardiographers on the basis of their identified kinematics parameters has in evaluated in [7]. The authors of [7] show that novices showed erratic movements which could be dangerous for the esophagus while experts showed smoothness and fluidity while handling the probe.

2.2.1 Virtual reality surgical simulators

A virtual reality surgical simulators consists of a dummy torso, a monitor screen and a probe which is identical to a standard TEE probe in terms of shape, articulation capabilities (flexion, rotation, angulation) and dimensions. The dummy TEE probe with an attached 3D tracking device is inserted inside the torso and manipulated by the trainee. The corresponding 2D echocardiographic view is shown on the monitor screen (Fig. 2.8). This image is identical to the one seen in real time surgery. A 3D virtual reality scene depicting the heart model and dummy TEE probe with the corresponding scan plan is shown in Fig. 2.8. Two companies offer a beating heart simulator model: Heartworks (Inventive Medical Ltd., London, United Kingdom) and Vimedix (CAE, Montreal, Quebec, Canada) [32].

Trainees learning with the help of such simulators, showing a side-by-side presentation of the heart plane, were able to understand the relationship between the

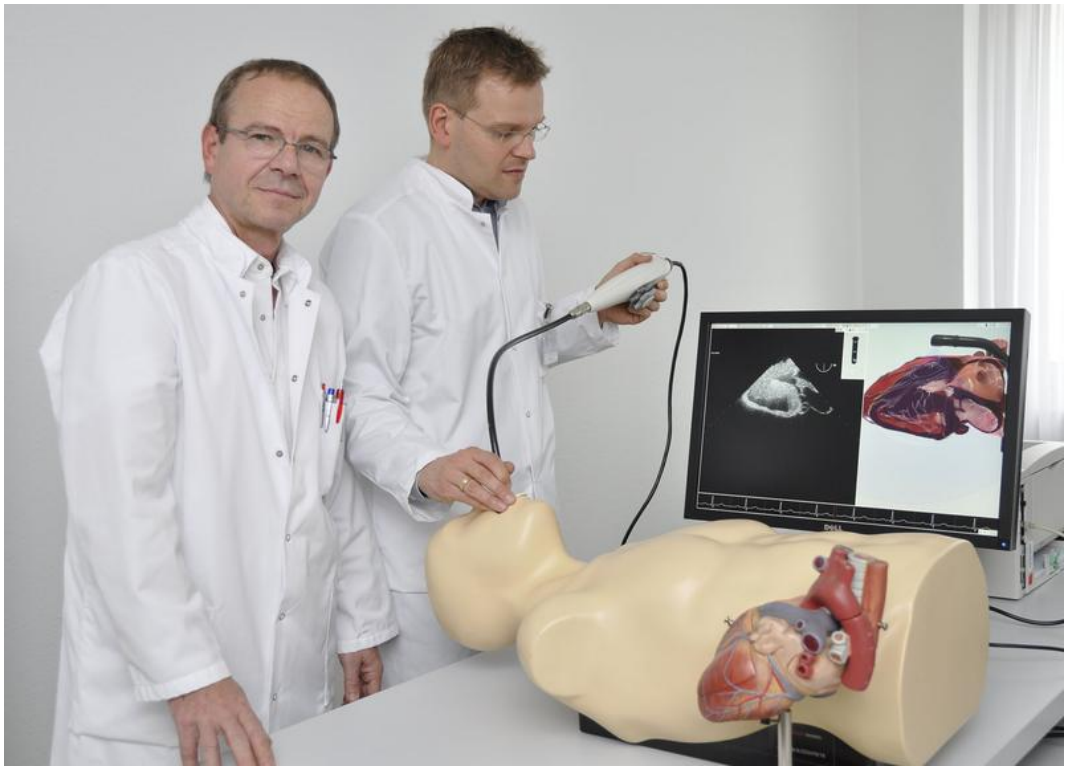


Figure 2.8: A practitioner carrying out TEE via HeartWorks virtual reality simulator. Image adapted from [5].

scan plan and the anatomy of the heart easily [33] [32]. The introduction of highly precise virtual reality surgical simulators has become a conventional approach for training new learners and for achieving proficiency in performing a TEE [32]. The problems mentioned in sec. 2.1.1 can be overcome with the advent of this maturing technology. Since, it replicates the surgical environment, a novice trainee can efficiently enhance his/her skills in a stress free environment.

Although a maturing technology, virtual reality surgical simulation stills needs a standardized artificially intelligent model to evaluate the images captured during TEE. In the upcoming section, the core idea of automating this process will be discussed.

2.3 Machine learning

Machine learning is concerned with the development of computer programs that have the ability to learn automatically by themselves and improve from their ex-

perience without being programmed explicitly. There are two approaches to train these programs (models), either by supervised learning or by unsupervised learning. In supervised learning, dataset with input feature *x* and output label *y* is provided. There are two techniques classification and regression which could be performed using supervised learning. In case of classification, y represents a definite set of classes and in case of regression, y is a vector with continuous values (i.e a continuous output is predicted in regression while a definite class is predicted in classification) [13]. In unsupervised learning, the data is processed without any labels and the model is trained to group or cluster data with similar characteristics [34]. With the help of machine learning, models could be built which allows the training of the models using pre-existing information and predict the desired outcome. Most recently, the advent of a new field called "Deep Learning" added further techniques to the machine learning domain: precisely, so-called "deep neural networks" (that is, neural networks composed of more than one hidden layer) discussed in the following sections.

2.4 Neural networks and deep learning approaches

The term "deep learning" refers to a subfield of machine learning, which is a part of artificial intelligence in computer science. Deep learning focuses on the computational models called as artificial neural networks which are based on or inspired by the structures and functions of the neural networks in a human brain. In simple terms, deep learning networks are the emulation of these biological neural networks which could be trained to capture highly non linear mappings between input and output modalities.

2.4.1 Neural networks

Neural networks are a type of learning algorithms with the interconnected artificial "neurons" inspired by biological neural networks (Fig. 2.10) [8]. The most basic neural network is called as perceptron (with any number of inputs and one output, as shown in Fig. 2.9). Neural networks with an input layer along with one or more hidden layers and one output layer are called as multi-layered perceptrons (MLP) .

Each layer can have any number of neurons which are connected to each other via links which have some weights W (W is a matrix comprising of columns w_k).

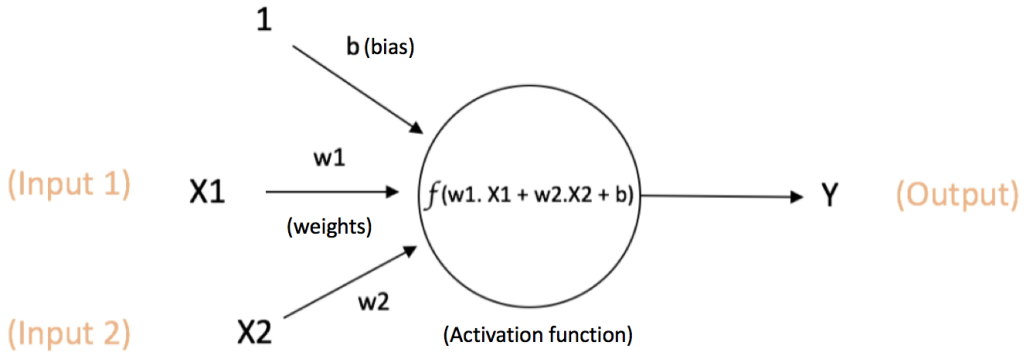


Figure 2.9: A perceptron. Image adapted from [8].

These neurons interact with each other to pass the information to higher layers and finally, the output layer predicts the output y. This output y is obtained after the linear combination of inputs (X, W), is passed through some activation function (a non linear function σ like sigmoid, Relu , tanh [10] [11]). Here X is an input vector

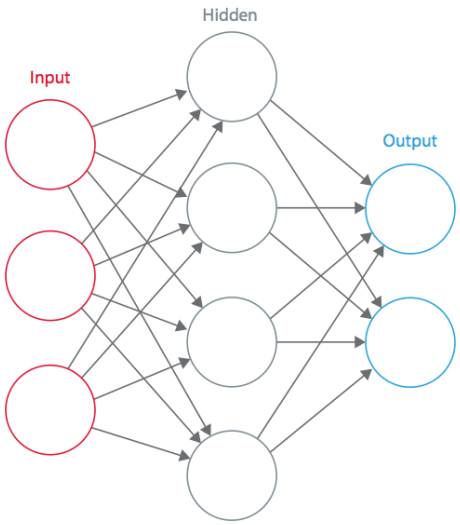


Figure 2.10: A sketch of a neural network. Image taken from [8].

as can be seen from fig. 2.9. Here b represents the bias which is added as a constant and W^T means transpose of W.

$$y = \sigma(W^T X + b)$$

The multi-layered perceptrons (MLP) have several of these transformations:

$$y = \sigma(W^T \sigma(W^T ... \sigma(W^T X + b) + b) + b)$$

Every connection between the neurons has some weight which is learned via gradient based optimization method called backpropogation (discussed in sec. 2.4.4). In the training phase, the network is fed with input X whose correct labels are already known (refereed to as ground truth). The input data travels through the network in a feed forward manner and finally passes through the last layer where the activation function is a softmax function (used only for classification tasks) to predict a label (output) for that input data.

$$Softmax = \frac{e^(W_i^T x + b_i)}{\sum_{k=1}^{K} e^(W_k^T x + b_k)} ,$$

where W_i is the weight vector leading to the output node associated with class i. This predicted output is then compared with the ground truth and the loss is evaluated using an appropriate loss function like mean squared error, cross entropy loss, etc. [35]. The gradient of this loss is backpropogated to update all the weights. This process represents one epoch and neural networks are trained with several epochs iteratively, till the loss is minimized and stops varying (Fig. 2.11) [8].

In 1950s, Rosenblatt introduced his famous perceptron model, this is how neural networks were first introduced [36]. Then in 1970s Multilayered Perceptron (MLP) models were introduced where a feed-forward neural network was created by stacking many perceptrons together to form a layer [37]. In the mean time, between 1960s and 1980s, the work on backpropogation algorithm (famously known algorithm now) was going on. Eventually, an efficient training of a multilayer per-

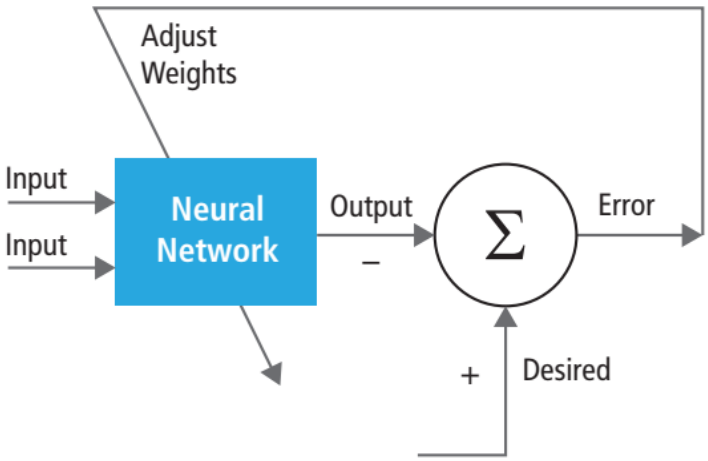


Figure 2.11: The workflow required for training a neural network. Image adapted from [8].

ceptron via backpropogation algorithm was demonstrated by Rumelhart, Hinton and Williams in 1986 [38]. During the same time, for the first time CNNs were introduced by Fukishima under the name Neocognitron [39] [40]. Currently, the research in the areas of neural networks and deep learning is at the front-line of machine learning.

2.4.2 Convolutional neural networks (CNN)

Convolution Neural Networks are a special kind of neural networks, in particular a type of a Deep Neural Network (DNN), neural networks composed of more than one hidden layer. They are the deep learning architectures which are capable of capturing non linear mappings between input and output. The input to a CNN network is a 3D matrix of an image with the first two dimensions representing as width and height of a RGB image and the third dimension corresponding to the number of feature maps (discussed shortly) learnt by the network. CNNs are majorly used for Image analysis and predictions and have successfully been applied to many computer vision problems. It has also been used in the field of medical science for the diagnosis of various diseases like lung cancer, breast cancer, skin cancer, etc. Recent advances of CNN in the field of computer vision, has prompted a surge of interest to apply similar set of techniques to medical images [41] [8].

The basic principle of a CNN consists in extracting increasing pieces of infor-

mation from an image. These initial small features of an image are then combined later (that is, in deeper layers) in the network. For instance, the first layer will try to detect some small features (like edges, lines, curves, corners etc), further layers will combine these features from previous layers to simpler shapes until, eventually, they will be combined to form templates of different objects. The final output is the weighted sum of all these templates. Convolutional neural networks consist of two major components (Fig. 2.12): (a) filters (also called feature detectors) and (b) feature maps.

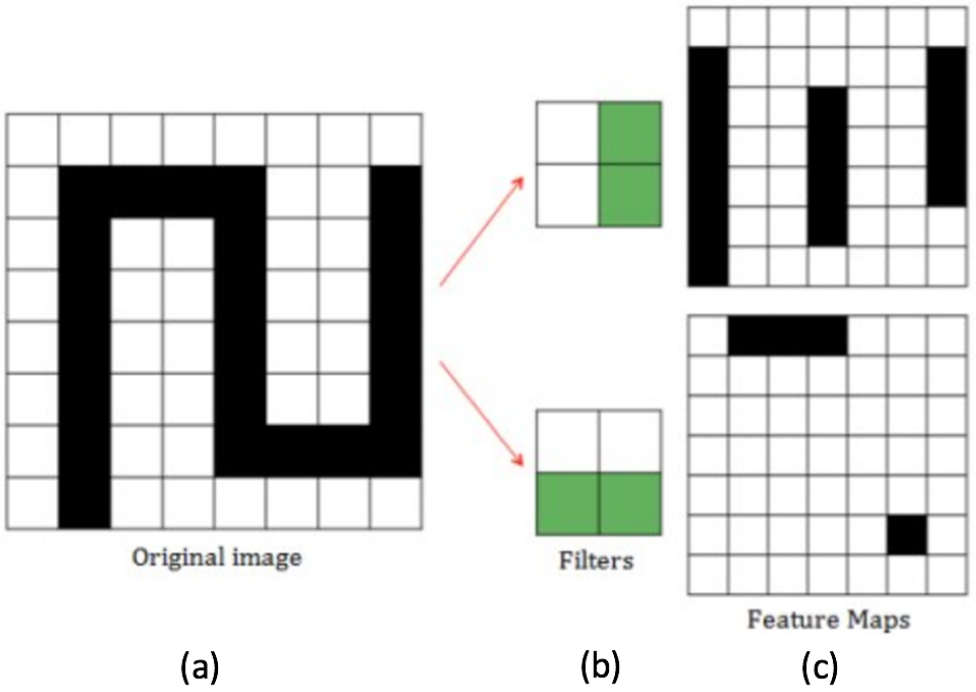


Figure 2.12: Filters after convolving over image generates respective feature maps. Image adapted from [9]

The two major components will be explained with the help of an example image (Fig. 2.12) [9]. In input image (Fig. 2.12 (a)) is given to a CNN and the task is to detect horizontal and vertical edges for which two filters (Fig. 2.12 (b)) are being used. The first filter detects the vertical edges and the second filter detects the horizontal edges. After convolving a filter over an image, a respective feature map is generated (Fig. 2.12 (c)) [10] [9]. In general, a filter is a square matrix of arbitrary size which is a prototype of the shape or the features we want to detect in

the original image.

2.4.3 Layers of CNN

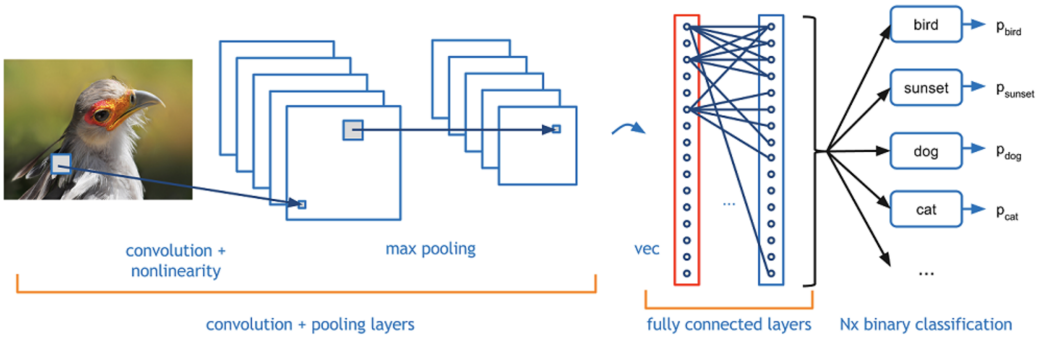


Figure 2.13: Systematic workflow of convolutional netural networks showing CNN layers. Image adapted from [10].

2.4.3.1 Convolutional layer

Convolutional layers are the main building blocks of CNN. This layer is the most important component of the network. Here each neuron is connected to only a subset of neurons of the previous layer instead being fully connected to them. The basic idea is highlighted in the following example. Assume we have an RGB image of size 32 x 32 x 3 (height x width x depth) which is cuboid in shape (Fig. 2.14 (b)). A filter (Fig. 2.14(a)) which detects the small level features such as lines, edges, curves etc. is convolved (scanned along the image) from the top left corner to the bottom left corner. The filter has the depth same as that of a cuboid block over which it is swept (for instance here it is, 5 x 5 x 3 i.e. depth is also 3 units). As the filter is sliding (convolving) through the image, each of the filter weights gets multiplied with the corresponding pixel value of the image. The sum of these multiplications give rise to the output (which is a single number) stored in the top left corner of the generated feature map (Fig. 2.15). This process is repeated until the filter has been convolved through the entire image [10][12][8][35].

The filter is convolved across the image with a **stride**, that is the amount with which the filter shifts along the image (Fig. 2.16). A stride of 1 results in a cuboid of size 28 x 28 x 1 (Fig. 2.14 (c)). The size of the length and breadth of the output

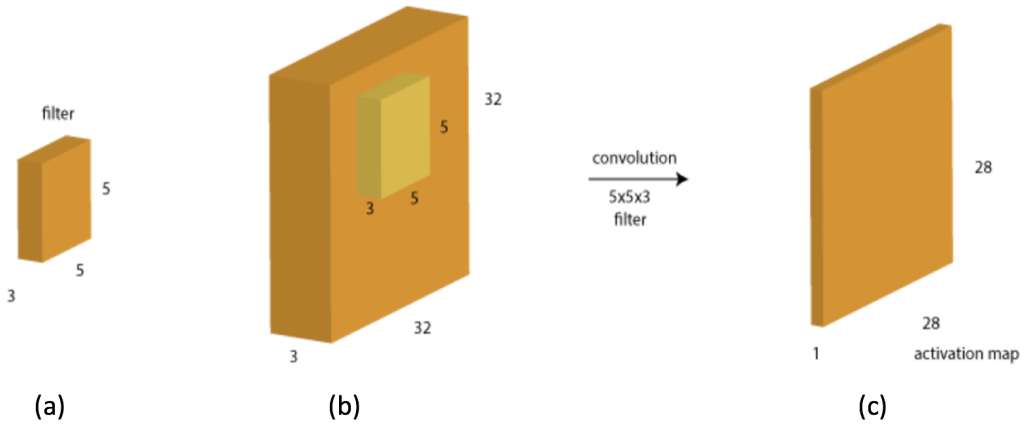


Figure 2.14: (a) Represents the filter; (b) Represents an input image over which filter is convolved; (c) The resultant activation map. Image adapted from [11]

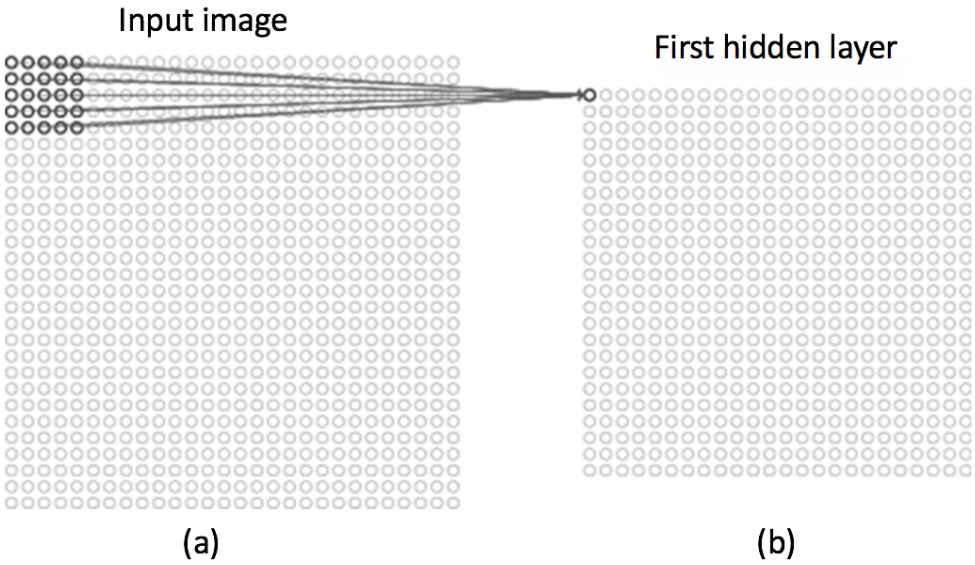


Figure 2.15: As the filter is sliding through the image, each of the filter weights gets multiplied with the corresponding pixel value of the image and sum of all these multiplications is stored in the top left corner of the generated feature map. Image adapted from [10].

is given by the formula

$$\text{output size} = (N - F)/S + 1,$$

where N corresponds to image size, F is the filter size and S is the stride. 5*5*3 +

1 = 76 values of the filter corresponds to 76 weights, that are the parameters collectively learned through back-propagation. The additional +1 represents a bias. Since the weights are shared throughout the image, the same filter is used throughout the image (though any number of such filters can be employed) [10][35].

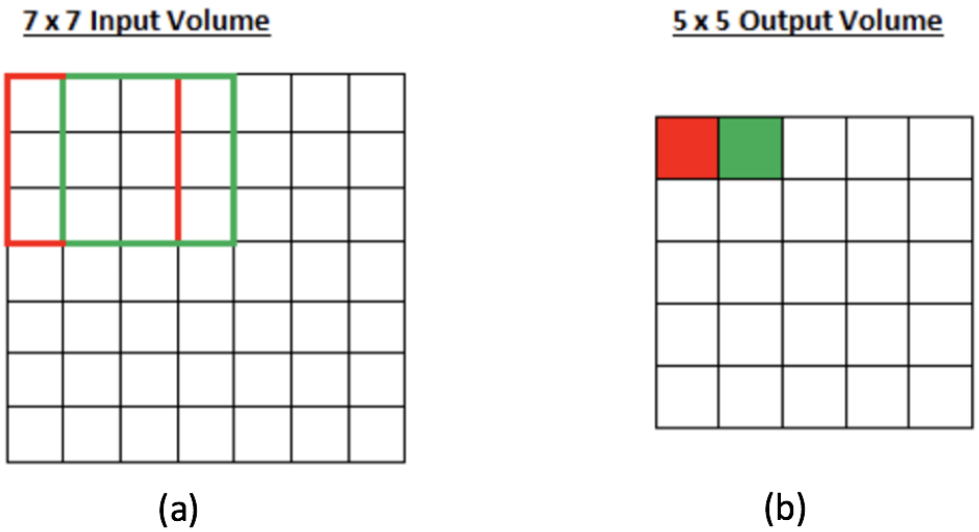


Figure 2.16: An example showing stride of 1. Image adapted from [12]

In a CNN network, there is not just one filter, it can have multiple filters according to the need of the user. The depth of the next layer is defined by the number of filters used in the previous layer. As can be seen from Fig. 2.17, 10 filters of size 7 x 7 x 3 are convolved over the image of size 32 x 32 x 3 giving rise to block of size 26 x 26 x 10 and so on.

As can be seen from Fig. 2.17, the size of the block (in terms of length and breadth) decreases as we progress deeper in the network. This is undesirable in case of deep neural networks as in the early layers we want to preserve the size so that lower level features could be extracted. To resolve this problem, **padding** which is a process of adding dummy pixels around the image is applied. Usually, zero padding of size $(F-1)/2$ (where F is the size of the filter) is done: this just means adding additional pixels of zero value along the border of the image as shown in Fig. 2.18 in order to restore its original size before the filter is convolved [11][10].

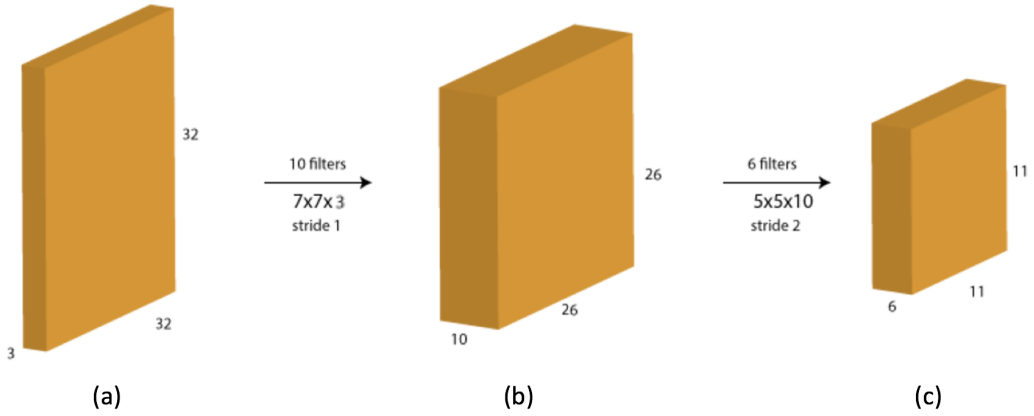


Figure 2.17: (a) Input image of size 32 x 32 x 3 ; (b) Block of size 26 x 26 x 10 obtained after convolving 10 filters of size 7 x 7 x 3 over an input image; (c) Block of size 11 x 11 x 6 obtained after convolving 6 filters of size 5 x 5 x 10. Image adapted from [11]

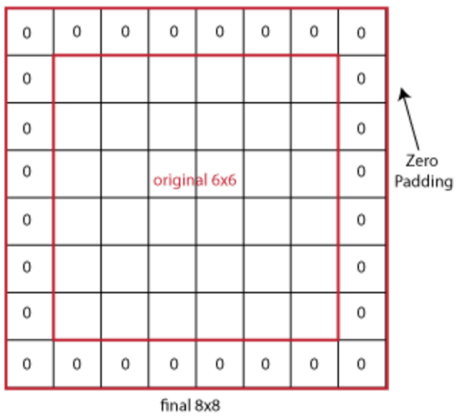


Figure 2.18: Image showing Zero Padding. Image adapted from [11]

2.4.3.2 Non linear layer

It is a convention for CNN networks [10] that after every convolutional layer, a RELU (Rectified Linear Unit) layer is applied to introduce non linearity to the network. An application of RELU basically applies $f(x) = max(0,x)$ to the input, thus changing all the negative activations to 0 (Fig. 2.19) [35].

Initially in CNNs, other activation functions such as sigmoid, or the hyperbolic tangent function (tanh) were used but nowadays, the RELU has emerged as a standard in CNNs [10]. This is due to the fact that RELU is capable of training a

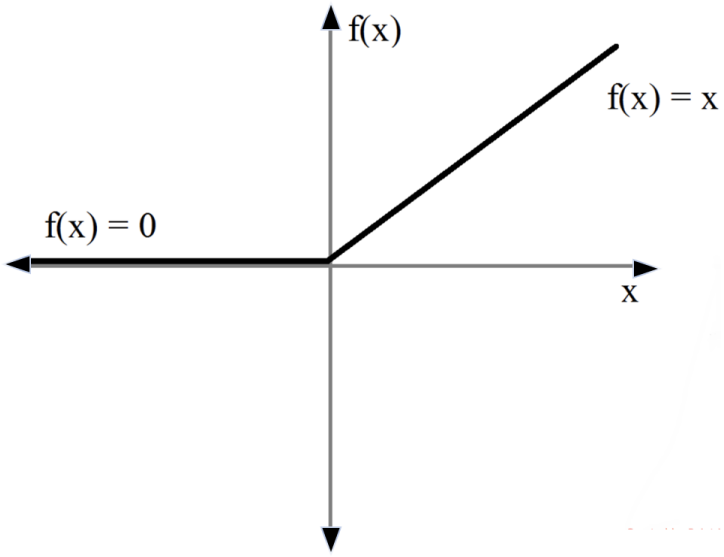


Figure 2.19: The non linear activation function RELU, $f(x) = max(0,x)$. Image adapted from [13]

network a lot faster than the previous activation functions.

2.4.3.3 Pooling layer

Pooling layer down-samples an input representation (image or hidden-layer output matrix) along the spatial dimensions i.e. width and height [42]. There are commonly used pooling techniques: *max pooling* and *average pooling*. Max pooling divides the image into subregions of size 2 x 2 and composes a feature map of reduced size by only keeping the maximum value out of each subregion. This results in a new size equal to the half of the one of the original image (Fig. 2.20).

Likewise, average pooling composes a feature map of halved size by only keeping the average value out of each subregion. Of the two approaches, max pooling is more prominently used [10] [35].

2.4.3.4 Fully connected layer

The *fully connected layers* are the last layers of a CNN network. Here, each pixel value is considered a single neuron just as in a general neural network. The number of neurons correspond to the total number of classes to be predicted [35]. The output of the fully connected layer is passed through softmax activation (discussed in sec. 2.4.1) to perform classification tasks.

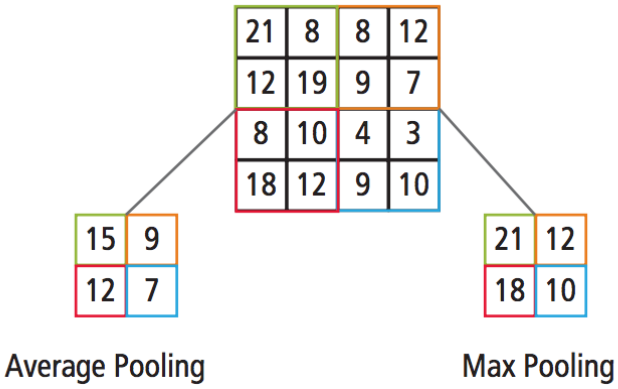


Figure 2.20: Image representing the two pooling techniques: average pooling and max pooling. Image adapted from [8]

2.4.4 Backpropagation

The training of a CNN network is done via **backpropagation**. Backpropagation involves four steps: forward pass, loss evaluation, backward pass and weight update. These four steps are performed for a fixed set of training images (in a batch job) until there is no more change in the weights of the filter (used as convergence criterion). Initially the weight vector W of the filter are initialized randomly in the forward pass. Since, the weights are random, we expect the output of the classifier to be uniformly random in the number of labels. After obtaining a new output from the classifier, we evaluate its loss L (defined as the Euclidean difference between the known training target and the obtained output vector) using the mean squared error (MSE) i.e. $L = MSE = \frac{1}{n}\sum_{(} target - output)^2$.

Overall, the loss can be computed as a function of the weight input vector. To reduce this loss, we thus perform a gradient descent on the loss (that is, $\frac{dL}{dW}$) to optimize the weight parameters. This is done at each layer where back-propagation in the backward pass is done to determine which weights contributed most to the loss (and hence to compute the gradient). After completing the backward pass, the weights are updated as usual in gradient descent using $W_{new} = W_{old} - \eta \frac{dL}{dW}$ where η is the *learning rate* (step size). [10] [43]

2.4.5 Deep learning architectures

There has been a lot of improvement since last decade in the field of convolutional neural networks and computer visions. Many different deep learning architectures have been introduced. In 1998, LeNet was introduced by LeCun et al. consisting of two convolutional layers [44]. This was a shallow network, then in 2012 the famous AlexNet was introduced by Krizhevsky et al. with five convolutional layers and three fully connected layers [19]. AlexNet employed dropout (discussed in section 3.4.1.1 to reduce overfitting in the fully connected layers and also won the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge), a computer vision challenge held annually which started in 2010 [45]. After 2012, further research in introducing deeper architectures started. In 2013, ZFNet was introduced which was an improvement to AlexNet [46]. Then in 2014 VGG was introduced with 16-19 variant layers and VGG19 won the 2014 ImageNet challenge for classification and localization. VGG is a popular network because of it's simplicity and depth [47]. One more 22 layer CNN architecture called GoogLeNet (also called InceptionNet) won 2014 ImageNet challenge for object detection with additional training data [48]. In 2015, He et al. introduced ResNet architecture which won 2015 ILSVRC consisted of so-called ResNet-blocks[49]. Since 2014, there was a saturation in the performance on ImageNet benchmark. Consequently, simple models such as AlexNet and VGG are the most popular ones in the field of medical science due to it's simplicity, performance and fewer memory requirements as compared to other CNN architectures [21].

2.4.6 Transfer learning and fine-tuning

Training of a CNN from scratch is a tedious and time consuming process as it requires large amount of annotated data which, in general, is difficult to obtain in the context of medical imaging where the expert annotation is expensive. In such situations, transfer learning technique is applied where existing pre-trained CNNs on millions of images (usually on natural images) are fine-tuned to make it compatible with medical images being used. There are mainly two scenarios in transfer learning:

• **Pre-trained CNN as a feature extractor:** In this strategy, a CNN pre-trained on large datasets (e.g. ImageNet is a dataset which has more than 15 million labeled high-resolution images with roughly 22,000 categories) is taken and it's last fully connected layer is removed (the output of this layer is the score obtained for defined number of classes for a different task). And then these extracted features are plugged into the existing image analysis pipeline for the new dataset.

• **Fine-tuning CNN:** In this strategy, the classifier on top of the CNN is replaced and retrained, and along with it fine-tuning of the weights of the pre-trained network via backpropagation (discussed in section 2.4.4) is also done. According to the problem under consideration, sometimes the entire layers of the ConvNet are fine-tunned and sometimes some of the starting layers are fixed and fine-tuning of only last few layers is done [50] [51].

Both the strategies have been used widely. Many papers have been published in which the authors use pre-trained CNN networks and fine-tune it according to their dataset. In 2014, Razavian et al. showed that using pre-trained CNN networks on ImageNet data as a feature extractor performed better as compared to the highly tuned state-of-the-art systems in the tasks related to computer vision [52]. Also In 2016, Menegola et al. compared fine-tuning of pre-trained networks with CNNs trained from scratch by doing some experiments and showed that fine-tuning outperformed given the small dataset (around 1000) of images of skin lesion [22].

2.4.7 Why CNNs are better?

CNN can have great advantages over previous machine learning algorithms in image-related areas as feature extraction in CNN is done automatically whereas in other machine learning algorithms (such as SVM, linear regression,etc.) features have to be extracted first and then fed explicitly to the models. The following are the advantages of CNN [8]:

•**Regression and classification with CNNs:** Regression and classification both can be performed effectively with CNNs on images. CNNs can be used

effectively as feature extractors from images as explained in section 2.4.6. Once, the features have been extracted, then these features can be used either to perform regression or classification based on the desired applications. 3D face reconstruction has been done in [53] using a CNN architecture that does regression. Alex Kendall et al. demonstrates that CNNs can be used to solve complicated regression problems using images as inputs through their famous architecture PosNet [54].

- **Memory Efficiency:** Since only the last layer is fully connected as the weights are shared for the entire image, CNNs have fewer memory requirements as compared to neural networks where every neuron is connected to each other [8].

- **Easier training and better performance:** If an equivalent network to CNN is built using standard neural network, it will have an enormous number of parameters which will require huge amount of computation power and hence will be difficult to train. Since in CNN, weights are shared, the number of parameters gets reduced drastically which is not only easier to train but also has better performance because it avoids overfitting by the introduction of pooling layers.

2.4.8 Applications of CNN in medical imaging

CNN has many applications in the fields closely related to Computer Vision such as medical imaging. Few of them are discussed below:

- **Image/exam classification :** This was the first area in medical image analysis where deep learning contributed majorly. In image/exam classification, the network is fed with an image (an exam) as an input and a diagnostic result (presence or absence of a disease) is predicted as an output (single variable). In such kind of classification tasks, the dataset is usually small where each diagnostic exam represents one sample of the dataset. Hence, for such kind of classification, transfer learning is the best solution as described in the section 2.4.6. CNNs are popularly used for exam classification task, all the 47

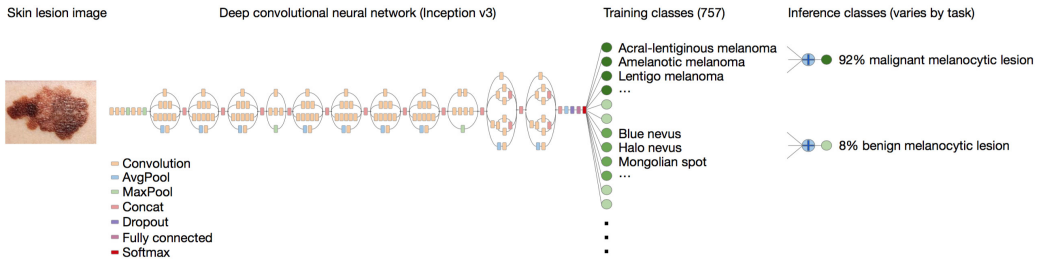


Figure 2.21: Image showing CNN classifying the images with the presence and absence of skin cancer. Image adapted from [14].

papers published in 2015, 2016, and 2017, 36 were using CNNs [21]. In-fact, CNNs pretrained on natural images have outperformed the accuracy of human experts and hence currently are used as a standard technique for diagnostic exam classification [21]. In 2016, Menegola et al. did detection of Melanoma (skin cancer) after training the dataset with 1000 images of skin lesions [22]. CNNs have been also been used for the detection of lung cancer and breast cancers effectively [55][56] [57].

- **Image Segmentation:** In image segmentation, each pixel in an image is assigned a label and similar characteristics are shared by pixels with same label [15]. Image segmentation is required for further analysis of an image by decomposing it into parts and thereby changing the representation of an image. In basic terms, by image segmentation the representation of an image is simplified which is easier to understand and analyze. It is used for locating boundaries

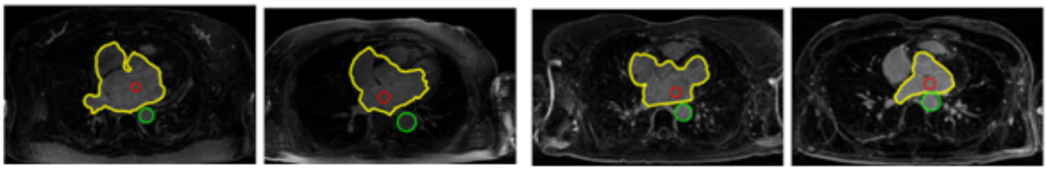


Figure 2.22: Image showing segmentation of Left Atrium, the yellow contour enclosing heart in a Cardiac Magnetic Resonance Imaging (MRI).[15]

i.e. lines, edges, curves, etc. of an object in an image. The type of image segmentation technique to be used on an image depends on its type and the problem to be solved. As can be seen in [15], image segmentation helps in

segmenting the boundaries of the heart (Fig. 2.22) in Cardiac Magnetic Resonance Imaging (MRI) to identify abnormalities within it's structure. It helps in accurately visualizing the cardiac movements, myocardial mass, etc. required for the diagnosis in MRI. Image adapted from[15] [21].

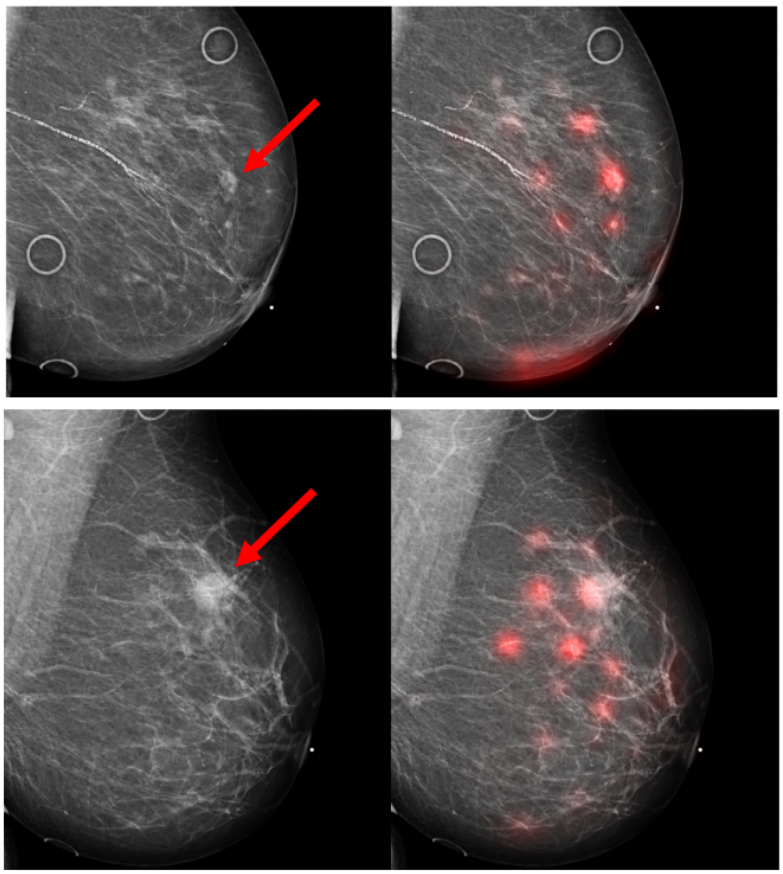


Figure 2.23: Image showing an example of object detection. The MRI images on the left side are of the breast of two patients with confirmed cancer (done by biopsy). The red arrows points the area which is proved to be malignant. The images on the right hand side are the predictions done by CNN network recognizing the suspicious findings (red region) with respect to the images on the left. Image adapted from [16]

• **Object Detection:** One of the most labor intensive and time consuming tasks for clinicians is to identify and localize the lesions or objects of interest in an image. In-fact, before performing segmentation, localization of organs and landmarks has been an important step in medical research. Researchers have been putting effort since long to design computer-aided detection systems which could detect lesions automatically and hence reduce the reading time

of human experts. CNN in object detection systems classifies each pixel and after classifying each pixel, some form of post processing is done to obtain the objects to be detected. In 1995, it was the first time an object detection system using CNN was introduced by lo et al [58]. Deep CNNs have also been used to detect breast cancer via object recognition. The clusters of microcalcifications were detected from the momographic screened images using CNNs [16] (Fig. 2.23).

Chapter 3

Methods and experiments

In this chapter the methods implemented and the experiments performed to automate the task of imaging skill assessment for TEE are described. To accomplish the aim of this thesis, the architecture shown is Fig. 3.1 was followed step by step. Initially, the dataset was acquired for each of the 10 planes of the heart from HeartWorks TEE simulator. Each of the heart image acquired was evaluated manually with two types of scores: criteria-based percentage score (0-100%) and a subjective general impression score (0-4). Then preprocessing of this data was done to obtain original heart images and further processing of these original images was done to obtain its four different versions. After that, two very well established deep learning architectures AlexNet and VGG16 were implemented. For each network, two regression models for regressing over each type of score and one classification model, to classify an image to any one of the ten defined planes of the heart, were constructed. There were three models constructed for each network (6 models in total) and finally, all the three models (for each network) were combined and the output was evaluated. These models were experimented with original images and also with its four different processed versions for detailed investigations. The training was done on GPU (graphical processing units which makes the computations faster), for 44 different combinations of models and images. First the methods and the approach followed for the problem under consideration have been discussed and then the experiments performed have been described briefly.

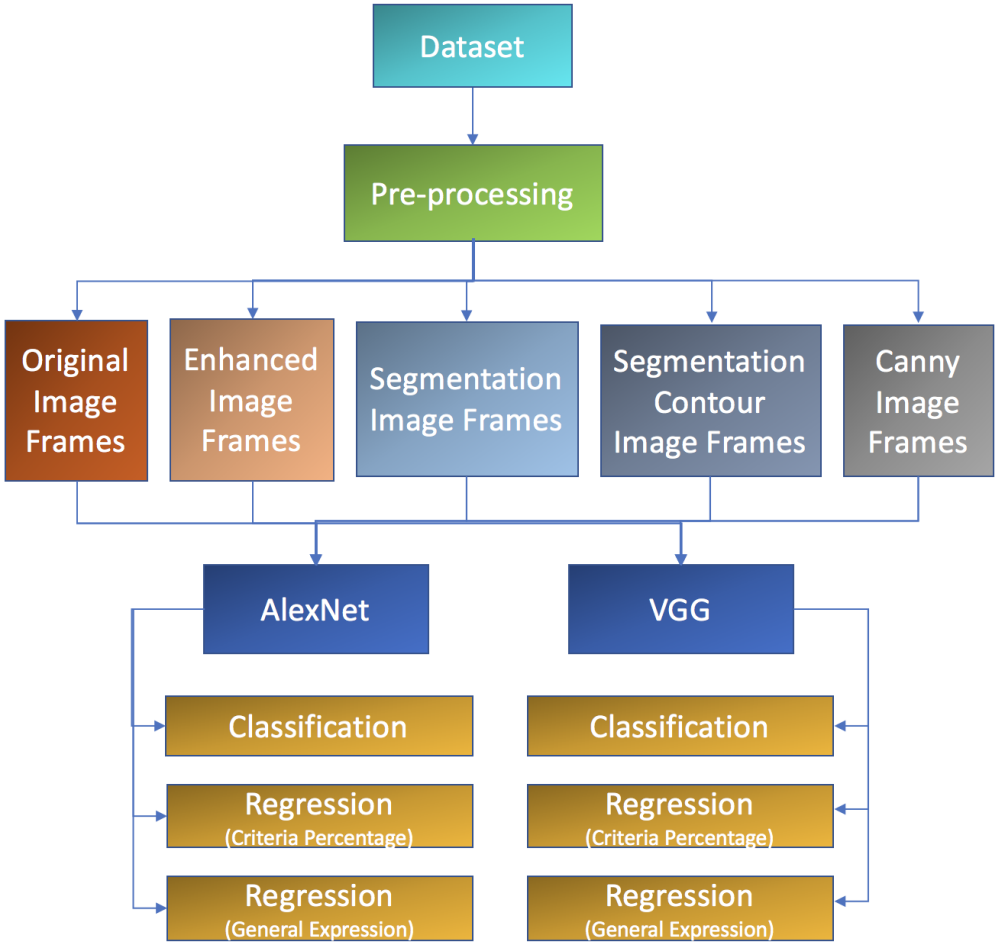


Figure 3.1: Image showing the architecture followed step by step to achieve the objective of the project.

3.1 Data Preprocessing

3.1.1 Data acquisition from HeartWorks VR simulator

The dataset utilized for performing experiments is obtained from HeartWorks TEE simulator (Inventive Medical, Ltd, London, UK). As shown in Fig. 3.3, HeartWorks TEE simulator constitutes an an upper-torso mannequin with mouth open so as to allow the insertion of the probe (Fig. 3.2). The probe employed with HeartWorks TEE simulator is similar to the standard TEE probe to get the required views of the heart (discussed in sec. 2.1.2) as shown in Fig. 3.4 [4].

For the assessment of the imaging skills of TEE which is the goal of the project, an experiment was designed where 38 participants were asked to obtain views of the

Figure 3.2: A trainee experimenting with HeartWorks TEE simulator. Image adapted from [5].

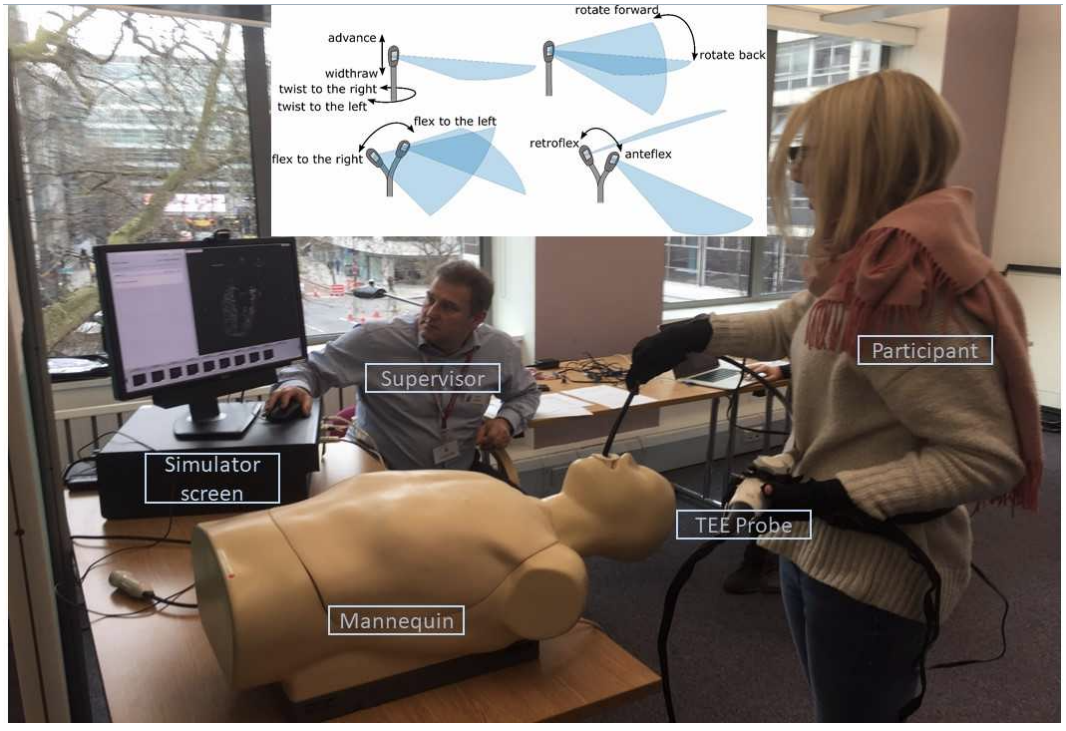


Figure 3.3: A participant performing TEE for the experiments done for the acquisition of data. Image adapted from [5].

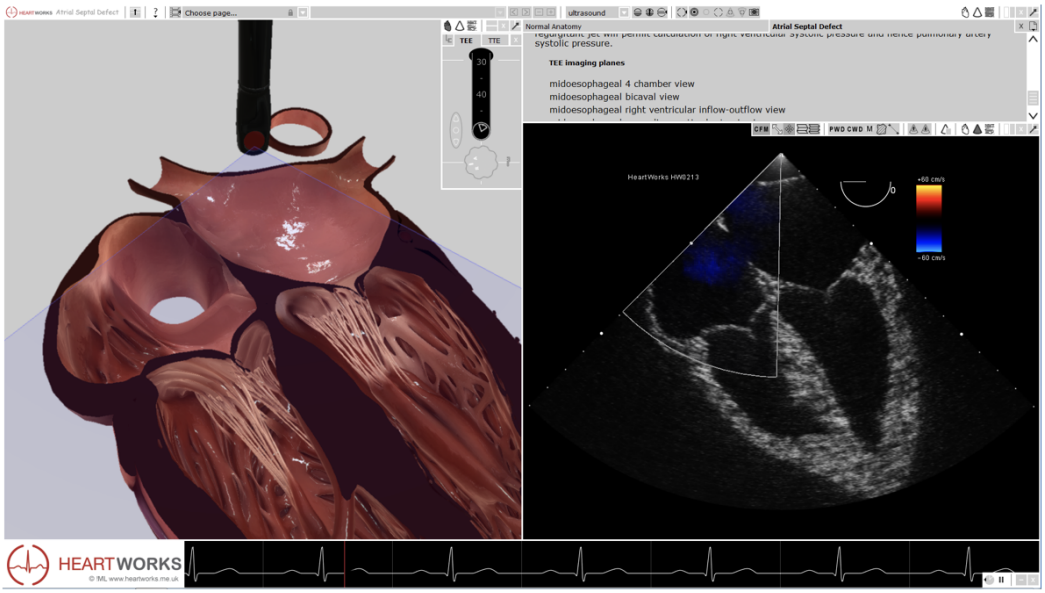


Figure 3.4: (a) Represents 3D rendering of the heart model; (b) Represents the corresponding simulated ultrasound view. Image adapted from [5].

Plane no.	Ultrasound image view
1	Mid-Esophageal 4-Chamber (centered at tricuspid valve) - ME4C (TV)
2	Mid-Esophageal 2-Chamber - ME2C
3	Mid-Esophageal Aortic Valve Short-Axis - ME AV SAX
4	Transgastric Mid-Short-Axis - TG mid SAX
5	Mid-Esophageal Right Ventricle inflow-outflow - ME RV inflow-outflow
6	Mid-Esophageal Aortic Valve Long-Axis - ME AV LAX
7	Transgastric 2-Chamber - TG2C
8	Mid-Esophageal 4-Chamber (centered at left ventricle) ME4C (LV)
9	Deep Transgastric Long-Axis - dTG LAX
10	Mid-Esophageal Mitral Commissural - ME MV commissural

Table 3.1: Sequence of the 10 ultrasound image planes used in the study

10 specific planes of the heart in a sequence as mentioned in table 3.1, with the help of the HeartWorks TEE simulator. These 10 specific views listed in table 3.1 are the subset of the 20 suggested views of the heart mentioned in the guidelines by American Society of Echocardiography (refer sec. 2.1.2 essential for a TEE examination [4]. The participants took these images under the tutelage of an expert anesthetist who guided them in taking the ultrasound images. The 38 participants

constituted 23 novice interventionalists (who didn't perform TEE exam more than 10 times) and 15 experts (accredited anesthetists who performed TEE exams more than 500 times). The novice participants were selected after testing their knowledge on the basic diagnostic aspects of TEE intervention and after making them practice on HeartWorks simulator twice during a one day introductory course on simulation. After the volunteers became familiar with the simulator, this experiment was performed. The views scanned by a participants were recored in the form of a short video (one second duration) for each of the 10 planes.

3.1.2 Manual data annotation

The assessment of aforementioned recorded videos from each participant (for each of the 10 planes) was done manually by three experts (accredited anesthetists) evaluator 1, evaluator 2 and evaluator 3 as shown in Fig. 3.5. For each plane to be scanned, there are certain criteria to be fulfilled as per the guidelines [4]. Each participant was marked on the basis of number of criteria he/she full-filled while scanning the plane. For example, plane number 1, 2, 3, 6, 8 and 9 required five criteria to be satisfied. Similarly, plane number 5 and 10 required 6 criteria and plane number 4 and 7 required 4 criteria to be fulfilled. Every plane has a different set of criteria to be satisfied, required to get the exact view of the heart plane [4][6]. Each

	View 1 ME 4C (TV)				
criterion number	**criterion**	**Evaluator 1**	**Evaluator 2**	**Evaluator 3**	**mean**
1	0-15 deg rotation	1	1	1	1.00
2	TV centred	0	0	0	0.00
3	no foreshortening	1	1	1	1.00
4	no LVOT in image	1	1	1	1.00
5	probe tip appropriately behind LA	0	0	0	0.00
6					
	criteria total	**3**	**3**	**3**	**3.00**
	criteria %age				**60.00**
	GEN IMPRESSION SCORE 0-4	**3**	**2**	**1**	**2.00**

Figure 3.5: A snippet from an excel sheet showing the marking done for plane number 1 for one of the participants.

examiner marks 1 for each criteria satisfied and 0 if it is not satisfied according to their assessment (Fig. 3.5). "Criteria total" sums up the score from each examiner

and finally the mean at the end is evaluated. The criteria percentage score is the is the quantized value i.e. for example (see fig. 3.5) a participant satisfied 3 criteria out of 5, so the criteria percentage score will be 60 % $\left(\frac{3}{5} * 100 = 60\%\right)$. This implies that the plane numbers which require 5 criteria to be satisfied can have criteria percentage score as 0%, 20%, 40%, 60%, 80%, 100% for satisfying 0, 1, 2, 3, 4, 5 criteria out of 5 respectively. **Here 20% corresponds to one criteria to be satisfied.** Finally the mean of these criteria percentage scores given by each examiner is evaluated and stored in the mean (Fig. 3.5), the mean score for the considered example will be 60% $\left(\frac{60+60+60}{3} = 60\%\right)$. This is the final criteria percentage score. The plane numbers with 6 and 4 criteria had quantized score too on the similar lines. The second type of score used for evaluation is the General Impression score which is given by the examiner just be looking at the scan and not checking the criteria. This score is between 0-4 and again the mean is evaluated from the scores obtained from the three examiners.

To prepare the labels for each of the ultrasound heart image, a CSV file (comma separated values) was generated (Fig. 3.6). This CSV file consist of five columns. They are described as:

• **Select_id:** Contains the name of the frames obtained mentioned in section 3.2.

	A	B	C	D	E
1	Select_id	Class	Criteria_Avg	Criteria_Pct	Gen_Impression
2	ah4_1.jpg	4	3	75	2.5
3	bc8_35.jpg	8	3.6667	73.333	3
4	dl7_25.jpg	7	4	100	3.6667
5	af5_43.jpg	5	5.6667	94.444	3.6667
6	cf5_37.jpg	5	5.3333	88.889	1.3333
7	be6_41.jpg	6	3	60	2.3333
8	ac3_1.jpg	3	2	40	2

Figure 3.6: A snippet of the CSV file generated.

• **Class:** This column represents the class for the image frame i.e. the corresponding plane number to which the image frame belongs to.

- **Criteria_Avg:** It is the average score of the criteria total from the excel sheet.
- **Criteria_Pct:** It is the average score of the criteria percentage score from the excel sheet.
- **Gen_Impression:** It is the average score of the general impression score from the excel sheet.

3.2 Video preprocessing

This section describes the preparation of the input image frames for the CNN. The dataset from one participant comprised of ten short one second videos for each of the 10 views of the heart i.e. for each participant, 10 videos were recorded. Since, there were 38 participants, there should be 380 videos. But, some of the videos for some planes weren't recorded correctly. Hence, 366 videos were there in total. These videos were transformed into images by converting each video into 44 frames (44 frames per second). These frames were then further processed to remove the HeartWorks logo and other two white patches on the right and the left side of the triangular region (Fig. 3.7). The image obtained after the preprocessing is shown in Fig. 3.8. There were total 16105 ($360 * 44$) images obtained from 366 videos, each having a resolution of 1000 * 1200. These images were stored in a folder and named as original image frames (as can be seen in the architecture shown in Fig. 3.1).

3.3 Image preprocessing

This section describes the further processing done on the original frames (i.e. images of the 10 views of the heart from each participant obtained in the former section) resulting into 4 different kinds of images for investigation purpose. These 4 different kinds of images were generated using different processing algorithms. These processing techniques were implemented to enhance the heart's structure and to remove the noise and specular highlights from the original frames. The first kind of image (Fig. 3.9) was obtained after using canny edge detector algorithm which is a technique to extract edges of an image. It extracts useful structural information and reduce the amount of data to be processed by large amounts [17].

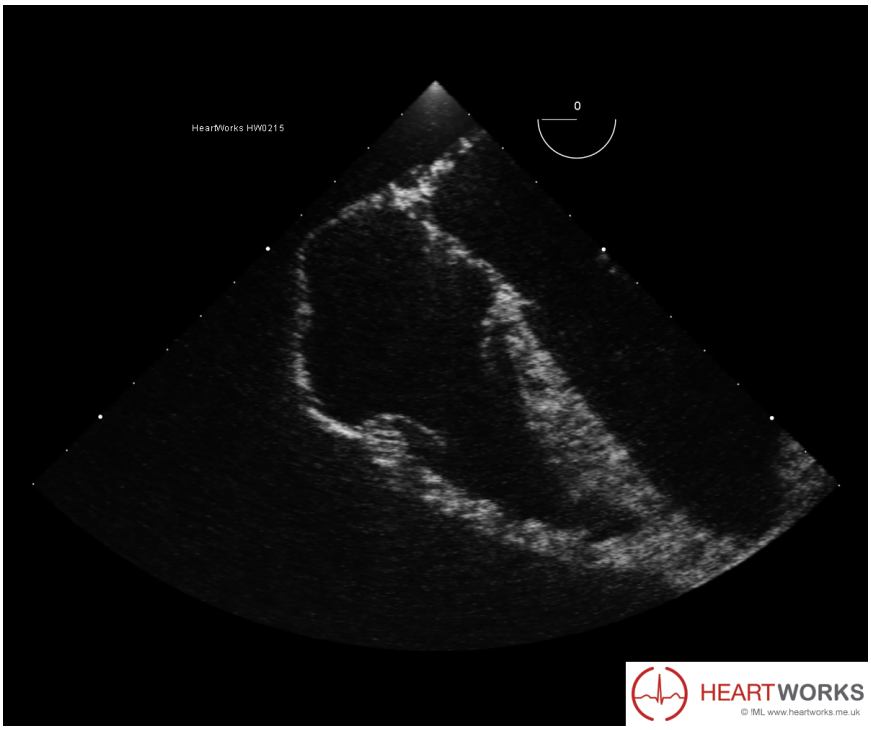


Figure 3.7: A sample frame obtained after processing a video (heart plane number 1) [5].

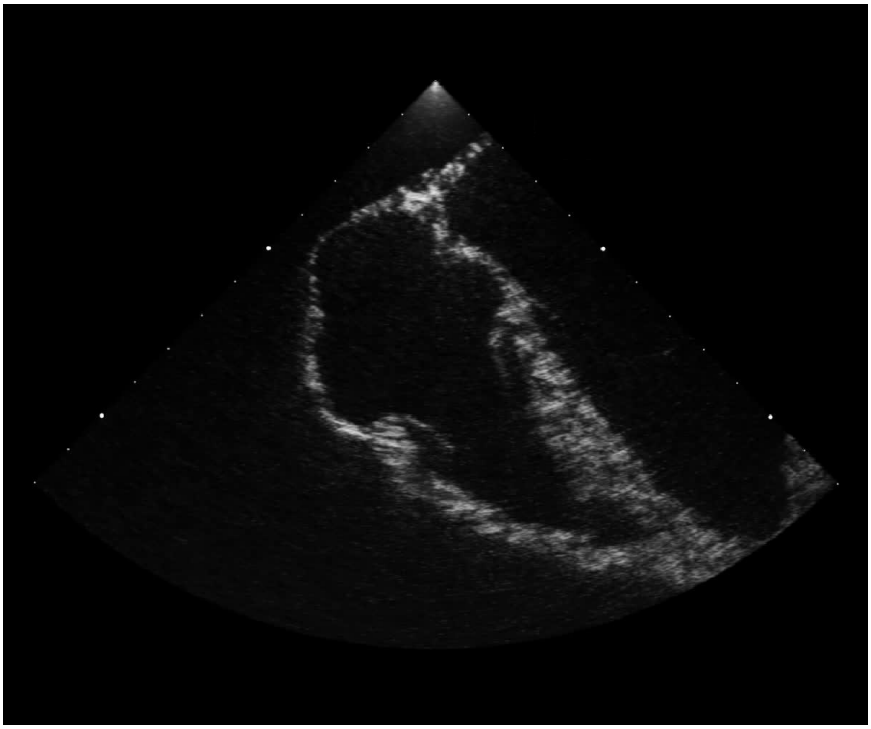


Figure 3.8: The image obtained after processing the frame shown in (original image frame).

The second kind of image (Fig. 3.10), enhanced image frame is obtained after suppressing the noise and enhancing just the cardiac structure with respect to

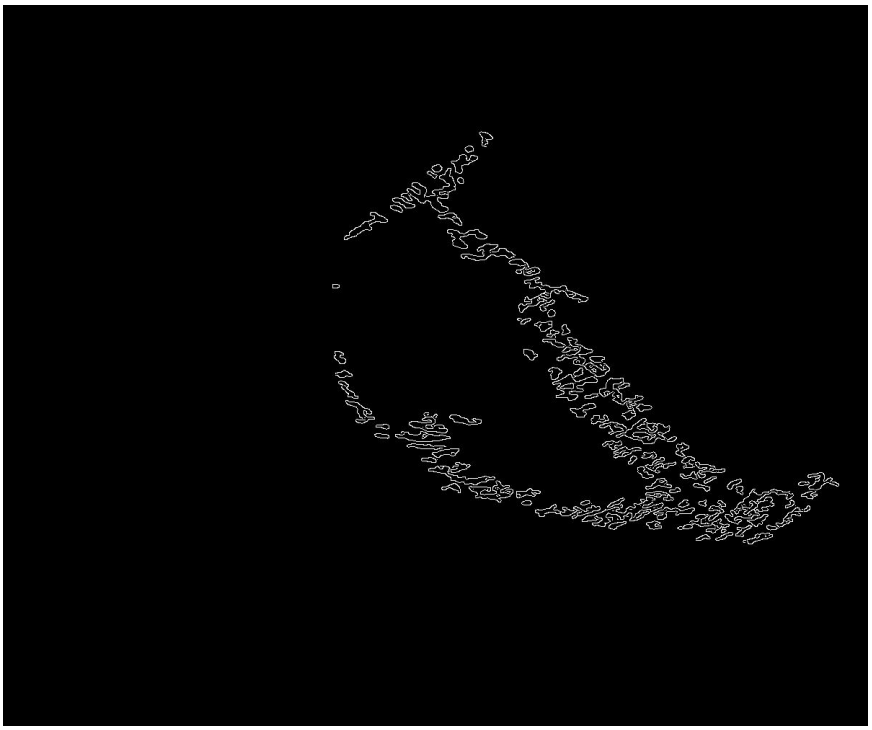


Figure 3.9: Canny image obtained using canny edge detector algorithm [17].

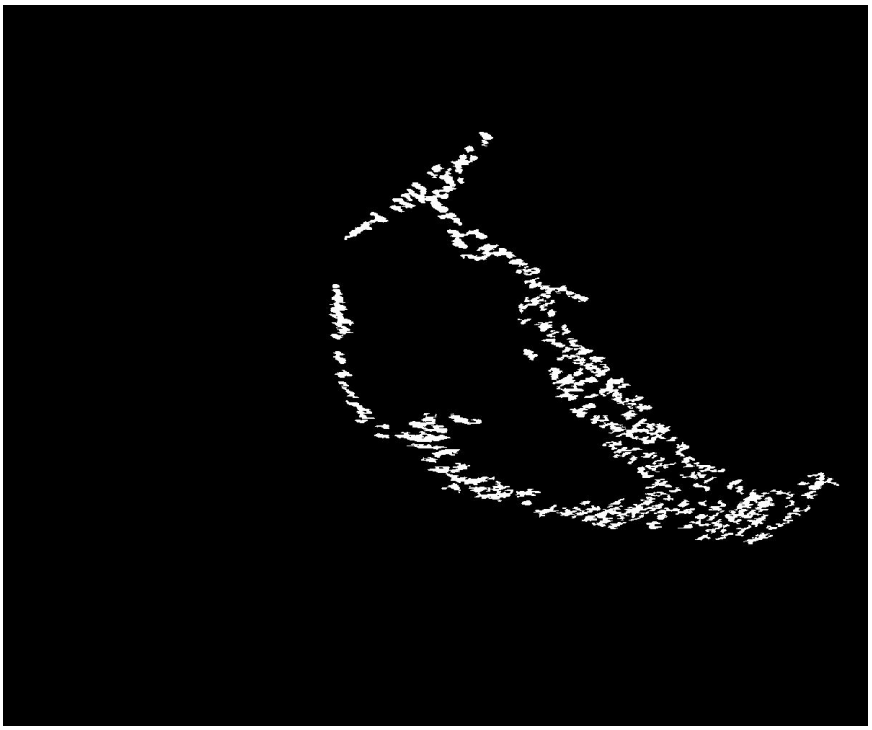


Figure 3.10: Enhanced image obtained after morphological processing with white hate filter and Otsu thresholding [17].

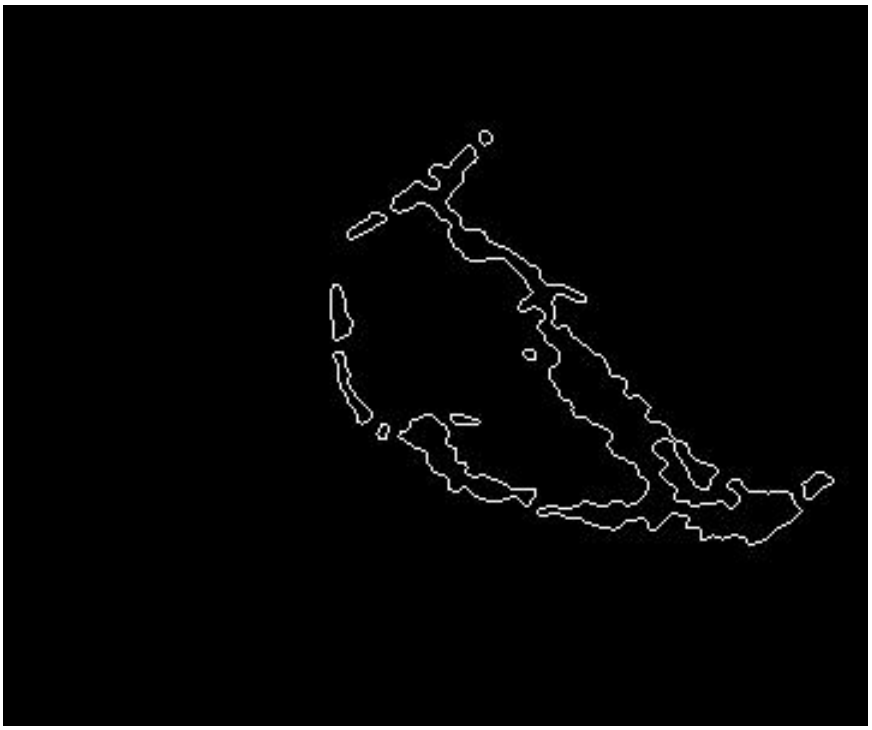


Figure 3.11: Segmentation contour image obtained using active contours segmentation algorithm [17].

Figure 3.12: Segmentation image after filling the above segmentation contour image [17].

the background. To eliminate the specular highlights and enhance the ultrasound plane, morphological processing with white-hat filter and Otsu thresholding was used. In the third type of image (Fig. 3.11), the segmentation of heart is done over the enhanced image. Due to thresholding, there were sharp edges around the border of the heart. Hence, using square-kernel Gaussian filter, the enhanced image was smoothed to produce segmentation image frames. The fourth type of image (Fig. 3.12), segmentation contour images was obtained by using a popular active contours segmentation algorithm introduced by Chan-Vese [59]. This image processing was done to for the purpose of detailed investigation. In the next section the architectures used to experiment with these images will be discussed briefly.

3.4 CNN architectures

After the preparation of data as described in the previous sections, in this section we will discuss the next step i.e. the implementation of two established deep learning architectures AlexNet and VGG16. AlexNet and VGG are established models for their performance on image classification tasks (discussed earlier in section 2.4.5) [19][47]. AlexNet won the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge), a computer vision challenge held annually which started in 2010 [45]. The VGG-16 model also won ILSVRC-2014 competition where it placed first in the in task 2a challenge. These models have been implemented a number of times in the field of medical imaging due to their simplicity and accuracy [60] [22]. In 2016, Hoo-chang Shin et al. used transfer learning approach to fine-tune AlexNet and other deep learning architetures to successfully detect thoraco-abdominallymph node (LN) and interstitial lung disease (ILD)[61]. VGG16 has also been used for the detection of skin cancer using transfer learning where it performed very well [60]. Since, Alexnet and VGG16 performed well with problems related to medical images, these two architectures became a choice for automating the task of skill assessment of heart views taken by the echocardiographers in the aforementioned experiment performed. Three models per network were built, they are as following:

- Classification of the heart scans over 10 planes of the heart: This model was built

to classify an image into one of the 10 planes of the heart.

- Regression on images with criteria_percentage score as label: This model was built to replicate the manual criteria percentage score. Regression with CNNs was done to get a continuous score for a given image. In simple terms it was done as instead of classification of the pixels of an input image, prediction of a single value close to the manual score was required.

- Regression on images with general_impression score as label: This model was built to replicate the manual general impression score. Regression with CNNs was done to get a continuous score for a given image. In simple terms it was done as instead of classification of the pixels of an input image, prediction of a single value close to the manual score was required.

These 3 models for each network (i.e. 6 models in total) were fine-tunned using transfer learning (discussed in section 2.4.6). After loading the pre-existing ImageNet weights corresponding to the networks, these three models were fine-tunned for all the layers. In case of **classification**, the code was modified to perform classification for 10 classes (original code classifies 1000 categories in both the networks) using softmax activation function after the last fully connected layer (shown in figure 3.14 and figure 3.15).

In case of **regression**, one more dense layer was added with number of classes as 1 (because we want the output as score and not the probabilities for 10 classes signifying the chances for that image to belong in a specific category which happens in classification). The output from the last fully connected layer was extracted (feature extraction explained in section 2.4.6) and was given to the dense layer added. The mean squared error (MSE): $Loss = MSE = \frac{1}{n}\sum_{(} target - output)^2$ was the loss function used during training for backpropogation (discussed in sec. 2.4.4). For regressing on general impression score, the label was taken as Gen_Impression and for regressing on criteria percentage score, Criteria_Pct was taken as label (see the snippet of the excel sheet shown in fig. 3.5).

An efficient deep learning framework called ”tensorflow” was used to run these

models a normal computer. There are other deep learning libraries too like caffee, theano, pytorch,etc. but tensorflow was chosen because of it's efficiency and flexibility. Tensorflow has an excellent visualizing module called tensorboard which makes the debugging easier and also allows to see the graphs for loss, accuracy, etc. at the run time. The AlexNet and VGG16 architectures used for the problem under consideration are discussed in detail in the upcoming sections.

3.4.1 AlexNet

There are two techniques implemented in AlexNet architecture:

3.4.1.1 Dropout

This approach was introduced by Hinton et al. to prevent overfitting in neural networks and CNNs during the training phase. Overfitting happens when the model

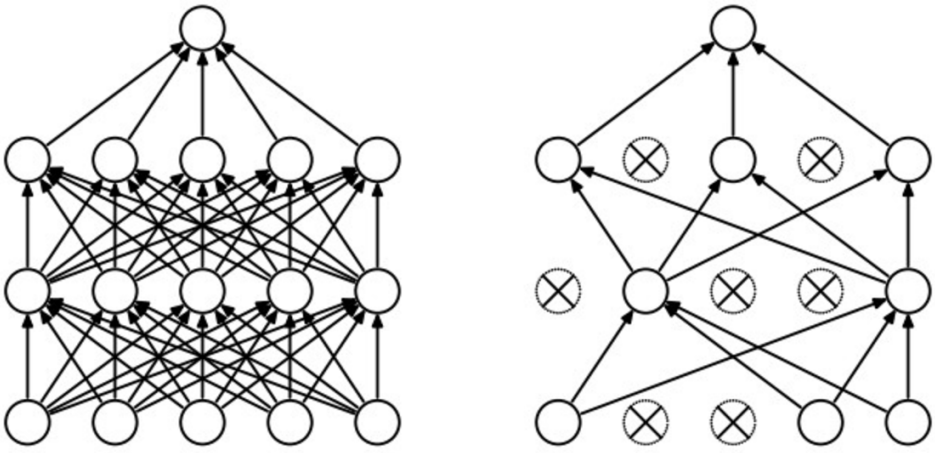


Figure 3.13: Overfitting prevention using dropout. Image adapted from [18]

learns the input data instead of generalizing the trend and hence performs poorly over the test data [62] [63]. In this approach (i.e. dropout approach), a neuron unit in the hidden layer is dropped temporarily from the network with some probability p (usually 0.5) randomly. This prevents these neurons to contribute to the network temporarily which enables the neural network to sample a different architecture (these architectures share same weights) each time (Fig. 3.13). This forces the neurons to learn more robust features and prevents them from co-adapting to

each other. This approach is similar to the method of ensembling different networks trained with the similar sets of parameters which is a successful way of reducing the errors incurred during test time (improving the performance thereby).

3.4.1.2 Local response normalization (LRN)

There is a concept in neurobiology called "lateral inhibition" which refers to the capacity of an excited neuron to subdue its neighbors. This neuron acts as a local maximum among it's neighbors which tends to increase contrast in that area thereby increasing the sensory perception. It is desired by a CNN to perform in the same way i.e. to dampen the responses of neurons which are uniformly large and make large activation more prominent within the neighborhood neurons and i.e. create higher contrast in activation map. CNN can achieve this lateral inhibition by implementing local response normalization (LRN). This layer is useful when dealing with RELU activation function because RELU has unbounded activations which needs to be normalized to detect high frequency features which have large response and diminish the others in neighborhood. The activity of activity of a neuron computed by applying kernel (filter) i at position (x,y) can be denoted by $a^i_{x,y}$. After applying the RELU nonlinearity, $b^i_{x,y}$ (response-normalized activity) can be given by the following expression:

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta},$$

where N represents the total numbers of kernels (filters) in the layers and the sum runs over n adjacent kernel maps at the same spatial position [19].

3.4.1.3 AlexNet architecture

Alexnet has a mandatory input size requirement of 227 x 227 x 3 [19]. The size of the input images was hence transformed from 1000 x 1200 x 3 to 189 x 227 x 3 to maintain the aspect ratio. And then these transformed images were padded with extra zeros so as to make it's size as 227 x 227 x 3. These images were fed in the form of a tensor to the CNN. Alexnet architecture contains eight layers: five convolution layers and three fully connected layers. It also contains max pooling

layers, dropout layer, local response normalization layers (explained in sec. 3.4.1.2 and RELU activation function (discussed in sec. 2.4.3 and sec. 3.4.1.1 of chapter 2). The first convolution layer uses 96 filters of size 11 x 11 X 3 with the stride of 4 resulting into 55 x 55 x 96 using the formula

$$\text{output size} = (N - F)/S + 1,$$

where N corresponds to image size, F is the filter size and S is the stride (mentioned in sec. 2.4.3) i.e. resulting into 96 feature maps of size 55 x 55.

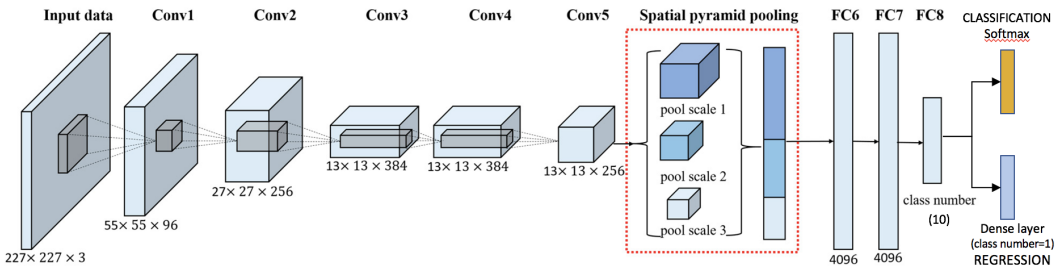


Figure 3.14: AlexNet Architecture used for performing classification and regression in the project. The last fully connected layer fc8 is modified according to the problem under consideration. Further, a softmax layer is added when performing classification to classify an image to any one of the ten defined planes of the heart and a dense layer with one neuron is added when performing regression using each of the manual scores. Original architecture adapted from [19].

The second layer is the max pooling layer (max pooling is done to reduce progressively the spatial size of the representation to reduce the amount of features and the computational complexity of the network) followed by a convolution layer which uses 256 filters of size 5 x 5 x 96 resulting into output of dimension 27 x 27 x 256. Local response normalization layers follows the first and the second convolutional layers after doing the max pooling. The next three layers follow the similar lines. The sixth layer is the fully connected layer where the input of size 13 x 13 x 256 is transformed into a vector of dimension 1 x 1 x 4096. The next fully connected layer follow the similar lines. A Dropout layer is added after both

fc6 and fc7 (fully connected layers). The last fully connected layer transforms the input vector to the dimensions of number of classes (10 in our case) i.e 1 x 1 x 10. For classification, this output was transfered to softmax activation function and for regression one more dense layer was added which gives the output as 1 x 1 x 1 as shown in the figure 3.14 [19].

3.4.2 VGG16

This subsection describes the detailed architecture of VGG16 used in this thesis. VGG16 has a mandatory input size requirement of 224 x 224 x 3 [47]. The size of the input images was hence transformed from 1000 x 1200 x 3 to 186 x 224 x 3 to maintain the aspect ratio. These transformed images were then padded with extra zeros so as to make it's size as 224 x 224 x 3. From each pixel of the image, the mean RGB value is subtracted, this is the only preprocessing done in the VGG16 network over the training set [47]. VGG16 is a sixteen layered architecture (more depth as compared to AlexNet). It contains 13 convolutional layers and 3 fully

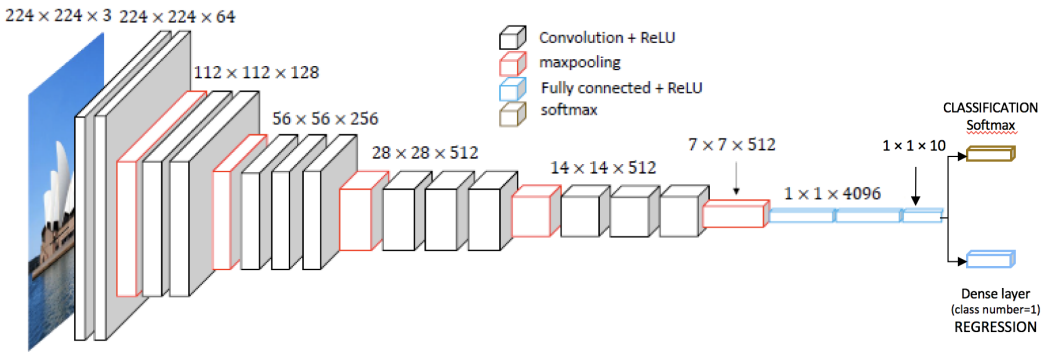


Figure 3.15: VGG16 Architecture used for performing classification and regression. The last fully connected layer fc8 is modified according to the problem under consideration. Further, a softmax layer is added when performing classification to classify an image to any one of the ten defined planes of the heart and a dense layer with one neuron is added when performing regression using each of the manual scores. Original architecture adapted from [20].

connect layers. The filters have the receptive field of size is 3 x 3 with the stride and padding both as 1 [47] so that their is no change in spatial resolution. It uses RELU as activation function and does max pooling over a 2 x 2 pixel window, with stride

of 2 to down-sample the input for easy computations (discussed in sec. 2.4.3.3). Out of the three fully connected layers, two layers have dimensions as 1 x 1 x 4096 and the last fully connected have dimension as 1 x 1 x 10 (same reason as that explained in AlexNet). On the similar lines as done in AlexNet, for classification, this output was transfered to softmax activation function and for regression one more dense layer was added which gives the output as 1 x 1 x 1 as shown in the figure 3.15 [47].

3.5 Dataset preparation

The data pre-processing has been done i.e. preparation of 16105 images from 366 videos to obtain 16105 images with the resolution of 1000 x 1200 and preparation of labels in the form of CSV file. Now, this data set was split on the basis of participants for training and testing data. Out of 38 participants, the data obtained from randomly chosen 6 participants (3 novice and 3 experts) was taken as the test set and the rest of the data obtained from the 32 participants was taken into training set. There was one more approach followed in splitting the dataset for training and testing, where the CSV file generated for training (see sec. 3.1.2) was shuffled and split into two parts. 80% of this file was kept for training and the rest 20% for testing.

3.6 Experiments performed

The training was done 44 times with 5 different kinds of images i.e. original image frames, canny image frames, enhanced image frames, segmentation image frames and segmentation contour image frames (discussed in sec. 3.3), 3 different models i.e. one classification model and two regression models (discussed in sec. 3.4) and 2 different networks i.e. AlexNet and VGG16. These models were trained and experimented using graphics processing unit (GPU).

- **Experiments done with AlexNet:** Each of the regression model built with Criteria_Pct and Gen_Impression as labels respectively was experimented with all the 5 different kinds of processed images as inputs i.e. each of the regression model was trained 5 times. The parameters chosen for the experimentation

were batch size= 128 and epoch=20 and the learning parameter chosen was .0001 for the performance reasons. Initially the training of models was done with no dropout probability, but the results obtained were not good, hence keepprob of 0.5 in the dropout layer was introduced and then the training was done. The classification was done only for original image frames with the same aforementioned parameters. The experiments performed using the dataset split over images (the another approach used) used the parameters as: batch size= 10, learning parameter = .0001 and epoch=10 since the loss graph on tensorboard started converging to minimum after 3rd epoch and started to over-fit when the number of epochs were increased.

- **Experiments done with VGG16:** Each of the regression model built with Criteria_Pct and Gen_Impression as labels respectively was experimented with all the 5 different kinds of processed images i.e. each of the regression model was trained 5 times). The models were experimented with the parameters as: batch size= 64 and epoch= 60 since the model was not performing good with less number of epochs since it is a dense network than AlexNet, it required more training. The learning parameter chosen was .0001 for performance reasons. The classification was done only for original image frames with the same aforementioned parameters. The experiments performed using the dataset split over images (the another approach used) used the parameters as: batch size= 40 and epoch=60 and learning parameter as .0001.

All the three models : the classification model and two regression models for Criteria_Pct and Gen_Impression scores respectively were combined and the output was evaluated. The metric used for evaluating the results of classification is the standard percentage accuracy and the metric used for regression models is root mean square error (RMSE): $RMSE = \sqrt{\frac{1}{n}\sum\left(actual\ manual\ score - predicted\ manual\ score\right)^2}$ where n is the number of test samples.

Chapter 4

Results and discussions

In this chapter, the results obtained after training 6 CNN models (3 models for each of the AlexNet and VGG16: one CNN model performing regression over criteria percentage score, second model performing regression over general impression score and the third model performing the classification of the heart image into the ten different planes of the heart) 44 times with 2 different networks (AlexNet and VGG16) and 5 different kinds of images: original, enhance, canny, segmentation contour and segmentation as mentioned in sec. 3.3 are shown in the following sections:

4.1 Loss graphs obtained after training the CNNs

This section discusses the training loss curves obtained from tensorboard for the regression performed on original frames with criteria percentage and general im-

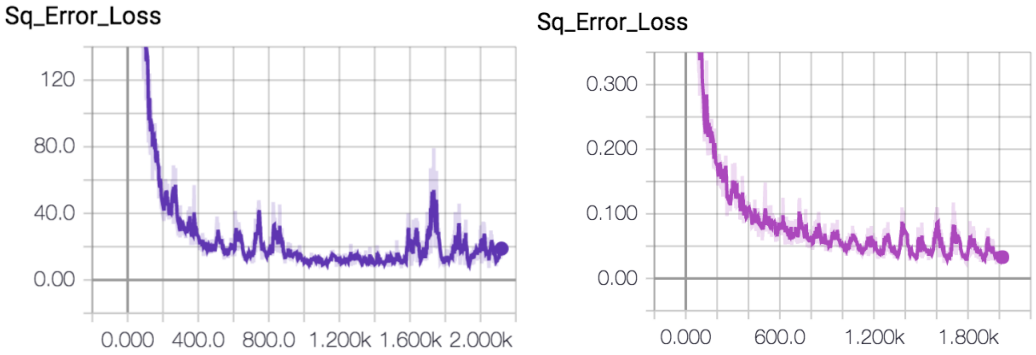


Figure 4.1: The first and second image represents the tensorboard loss graphs for the regression performed on original images (using AlexNet) with criteria percentage score and general impression score as labels respectively.

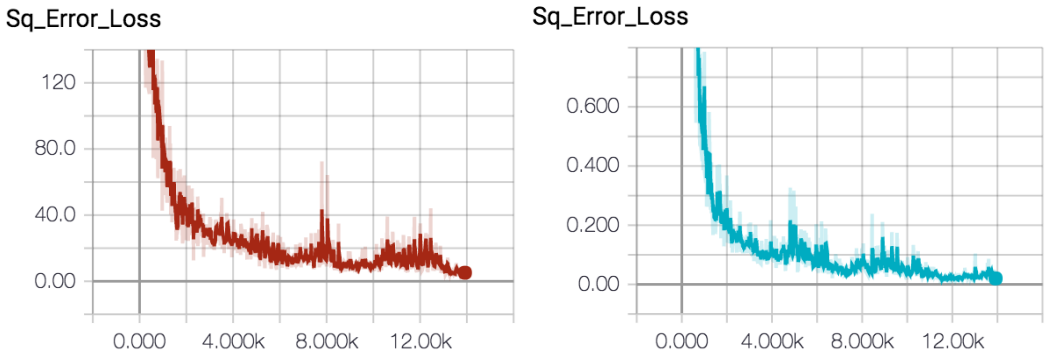


Figure 4.2: The first and the second image represents the tensorboard loss graphs for the regression performed on original images (using VGG16) with criteria percentage score and general impression score as labels respectively.

pression as labels using both networks (shown in figures 4.1 and 4.2). The X axis represents the total number of batch iterations i.e. $batchsize * epochs$ and Y axis represents the mean squared loss. All of these loss graphs are converging to minimum after showing some oscillating behavior. The loss graphs for regression performed on other type of image frames were similar to the ones shown here with the loss converging between the range 0- 15. The root mean squared loss for criteria percentage as label is for the range 0-100 while it is from 0-4 for general impression score as label. The graphs for VGG16 are converging to zero while in case of Alexnet the loss graphs are converging to the value relatively greater than zero.

4.2 Results obtained from different processed versions of original images

The results obtained after performing regression with criteria percentage score and general impression score using AlexNet and VGG16 for the different kinds of input images are listed in the tables 4.1 and 4.2 respectively. The first column in these tables represents the type of image frame used for training a CNN. The second and the fourth column lists the RMSE (root mean square values) obtained on the test set after regressing over criteria percentage score and general impression score as labels respectively using the corresponding image frames. The third column represents the RMSE value obtained after quantizing the criteria percentage scores [i.e. obtaining

the nearest value of the criteria percentage score (predicted by a CNN) with respect to all the quantized values possible for criteria percentage score (as explained in sec. 3.1.2) depending on the class to which the image frame belongs to]. In the following subsections, the RMSE values will be compared based on datasets:

Results obtained from AlexNet have been listed in table 4.1. RMSE values for criteria percentage is the lowest for original images (16.8084), enhanced images have similar results as original frames, the error is relatively more for segmentation frames, followed by segmentation contour frames and highest for canny images (19.1033). For general impression score, the RMSE values for original, enhanced, segmentation and segmentation contour is almost similar while it is relatively higher for the canny images.

Type of images	**Criteria percentage score (RMSE)**	**Quantized criteria percentage score (RMSE)**	**General impression score(RMSE)**
Original	16.8084	16.7154	0.7823
Enhanced	16.9081	16.8491	0.7875
Segmentation	17.5478	17.3529	0.7937
Segmentation contour	18.6009	18.4496	0.7762
Canny	19.1033	19.0120	0.8772

Table 4.1: The results obtained from AlexNet. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression.

Type of images	**Criteria percentage score (RMSE)**	**Quantized criteria percentage score (RMSE)**	**General impression score(RMSE)**
Original	9.6477	9.4895	0.5347
Enhanced	9.7487	9.5223	0.5678
Segmentation	10.1483	10.0496	0.4656
Segmentation contour	11.7138	11.6355	0.4004
Canny	12.5162	12.2785	0.5418

Table 4.2: The results obtained from VGG16. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression.

Results obtained from VGG16 have been listed in table 4.2. Following the trend

earlier, here also RMSE values for criteria percentage score is lowest for original frames (9.6477) and highest for canny images (12.5162). The error for enhanced images is similar to the original images while segmentation and segmentation contours are have relatively more error. The RMSE values for general impression score is almost similar for original, enhance and canny images while it is little bit low for segmentation and segmentation contour images. The classification model for AlexNet and VGG16 gave similar accuracies i.e. 98.021%. and 98.332% respectively .

4.3 Comparison of scoring techniques

In this section, the comparison of the results using criteria percentage score and general impression score as labels is done. This section also discusses the effect over RMSE values after quantizing the criteria percentage score predicted from CNNs. The figures 4.3 and 4.4 represents the error in percentage for each type of the image frame for AlexNet and VGG16 using dataset 2 respectively. The RMSE

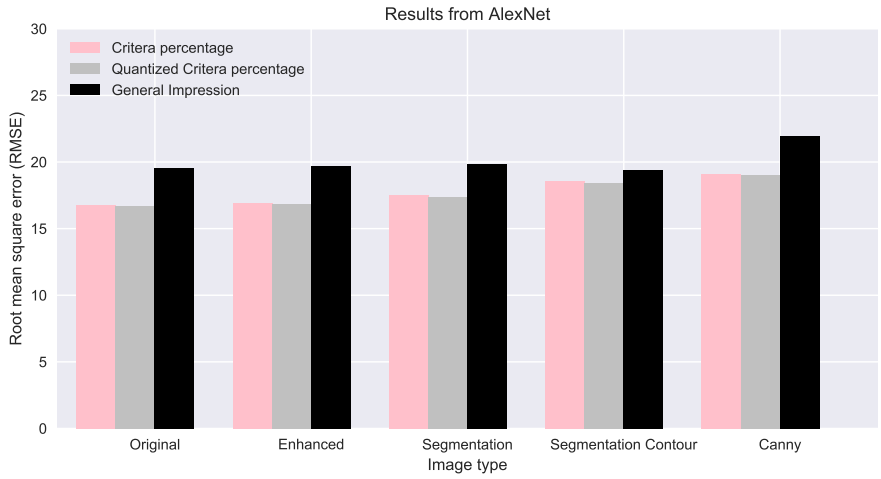


Figure 4.3: Results from AlexNet showing that the performance of regression over criteria percentage score is better than regression over general impression score. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively.

for general impression score is relatively higher as compared to criteria percentage and quantized criteria percentage for all the types of images in case of AlexNet.

In case of VGG16, a similar trend is shown for original, enhanced, segmentation and segmentation contour images while breaking the trend, segmentation contour images gives less RMSE for general impression score. The RMSE error obtained from VGG16 for general impression score is relatively lower for all the frames as compared to the one obtained from AlexNet (refer table 4.1 and table 4.2).

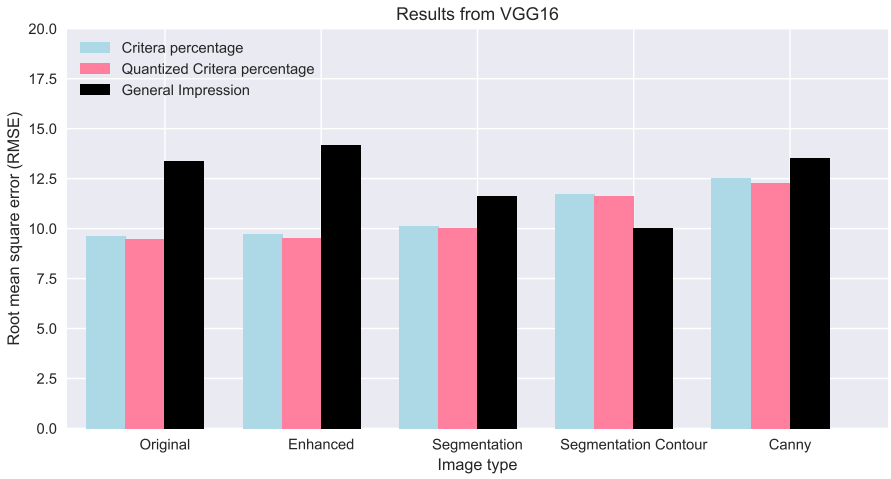


Figure 4.4: Results from VGG16 showing that the performance of regression over criteria percentage score is better than regression over general impression score for all the frames except segmentation contour images. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively.

The Quantized criteria percentage score is giving relatively less RMSE values as compared to the RMSE values for criteria percentage score for both the networks since after quantization the predicted values will become closer to the ground truth (as explained in section 3.1.2) thereby reducing the error. The estimation of error with criteria percentage score as label was lesser as compared to regression over general impression score. This is quiet logical thing to happen because the criteria percentage score is more accurate and standard since it is given on the basis of a number of minutely defined criteria, required to be satisfied while scanning the heart along the 10 planes while the general impression score is given by an expert just by looking at the scan and not checking the criteria.

4.4 Comparison of different networks

In this section, the networks based on the results obtained will be compared. The Root Mean Square Error (RMSE) for predicting criteria percentage score was in the range 16.80-19.10 % for AlexNet and 9.64-12.51% for VGG16 . The RMSE for predicting general impression score was in range 0.77-0.87 for AlexNet and 0.40-0.56 for VGG16. . For regression using criteria percentage as label, VGG16 is performing better than AlexNet by by 7.1607% and for general impression it is performing better by 6.19 % (after changing the scale of general impression from 0-4 to 0-100), because it is much dense network with 8 more convolutional layers. Hence, CNN extracts more features in case of VGG16 and performs better. For each of the heart plane there are certain criteria associated to obtain the perfect view for that plane. For example, Mid-Esophageal 4-Chamber (centered at tricuspid valve) - ME4C (TV) (plane number 1) requires Tricuspid Valve to be in the center [4]. Mid-Esophageal Aortic Valve Short-Axis - ME AV SAX (plane number 3) requires Aortic Valve of the heart to be in the center of the screen and so on [4]. Features from such criteria based shapes and templates are better extracted in VGG16 and hence prediction of criteria based score is closer to the manual scores. In case of classification of an heart image into 10 planes of the heart, both the networks are giving almost similar results, though VGG16 is still performing a bit better by 0.311%. The accuracies in case of classification for AlexNet and VGG16 are 98.021%. and 98.332% respectively.

4.5 Results obtained from the shuffled dataset

There was one more approach followed where the dataset generated was shuffled and split into 80% training set and 20% test set. This data was shuffled on the basis of images and not the participants and was giving the results on the similar lines as the aforementioned results. The RMSE values were relatively a bit better because the frames obtained from the video were almost similar for many videos and hence while splitting the data into training and testing, it might have been possible that the similar frames were present in both training and testing data resulting into lower

RMSE over test set. While, in case of the dataset which was split on the basis of participants, the training and testing data were completely different as none of the frames from the same video were in both the training and the testing data. Hence, the RMSE values for all the image types in case of dataset split on the basis of participants were relatively higher. The classification model for AlexNet and VGG16 using the dataset shuffled over images gave similar accuracies i.e. 99.032%. and 99.561% respectively.

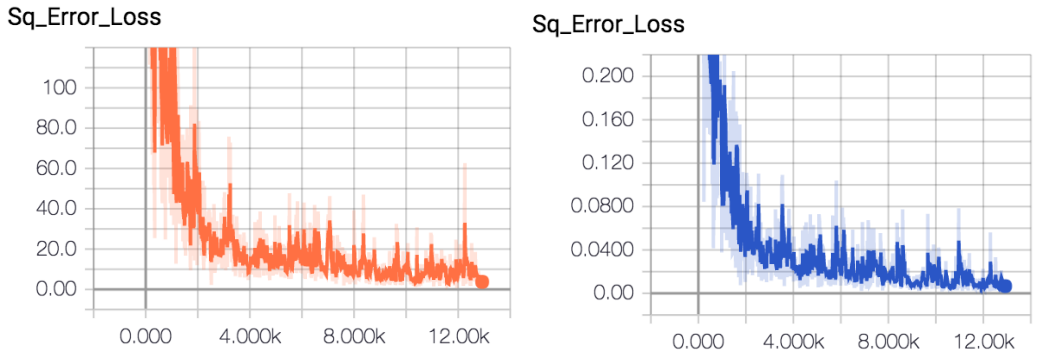


Figure 4.5: The first and the second image represents the tensorboard loss graphs for the regression performed on original images (using AlexNet) with criteria percentage score and general impression score as labels using the dataset shuffled over images respectively.

Figure 4.6: The first and the second image represents the tensorboard loss graphs for the regression performed on original images (using VGG16) with criteria percentage score and general impression score as labels using the dataset shuffled over images respectively.

Type of images	Criteria percentage score (RMSE)	Quantized criteria percentage score (RMSE)	General impression score(RMSE)
Original	3.3728	2.9037	0.1409
Enhanced	4.3733	4.0832	0.1858
Segmentation	4.5839	4.2788	0.2181
Segmentation contour	5.9882	5.6754	0.2052
Canny	5.6654	5.2877	0.2574

Table 4.3: The results obtained from AlexNet using using the dataset shuffled over images. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression.

Type of images	Criteria percentage score (RMSE)	Quantized criteria percentage score (RMSE)	General impression score(RMSE)
Original	1.7101	1.6398	0.1829
Enhanced	3.0539	2.8066	0.2432
Segmentation	3.5481	3.2115	0.2596
Segmentation contour	3.5802	3.2389	0.2052
Canny	3.2970	3.1068	0.28128

Table 4.4: The results obtained from VGG16 using using the dataset shuffled over images. The range of the values for criteria percentage and quantized criteria percentage is 0-100 and it is from 0-4 for general impression.

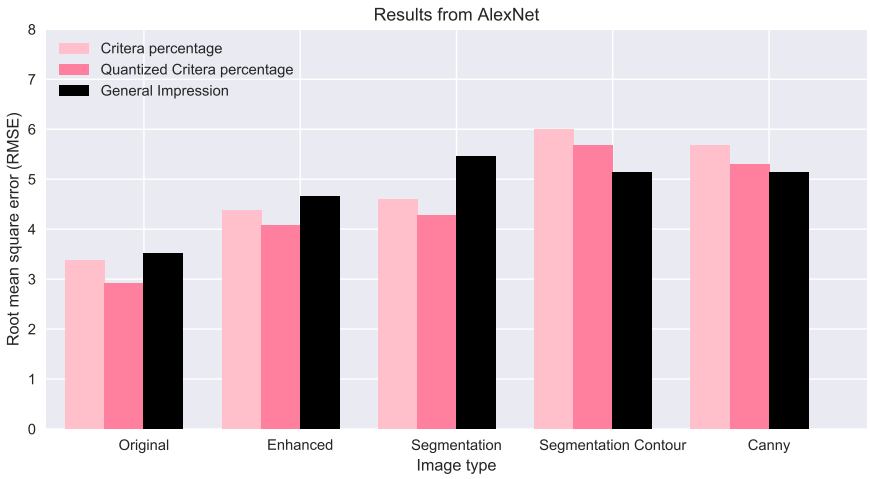


Figure 4.7: Results from AlexNet using the dataset shuffled over images. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively.

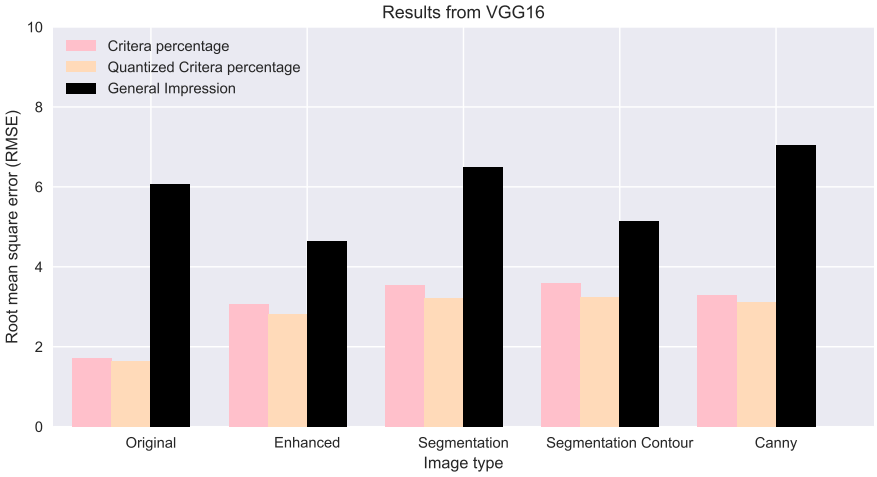


Figure 4.8: Results from VGG16 using the dataset shuffled over images. Here the general impression score is normalized i.e. the scale is changed from 0-4 to 0-100 for comparison. These bar graphs represents the error in percentage for each type of the image respectively. The performance of regression over criteria percentage score is better than regression over general impression score.

Chapter 5

Conclusions and future work

5.1 Summary of the work

The objective of this project was to build an automatic system which could replicate the manual scoring done by the experts,for the heart images obtained while carrying out a heart diagnostic test called Transoesophageal Echocardiography (TEE). To accomplish this goal, manual assessment process was carried out first. Experiments were done to acquire the data from HeartWorks TEE simulator where 38 participants (23 novice and 15 experts) carried out TEE tests individually. Each participant was required to take ten suggested views of the heart along ten different planes. Hence, there were ten types of heart images obtained by each participant. Each type of heart image was manually evaluated by the experts for two different scores namely, criteria percentage score and general impression score. For Criteria percentage score, several criteria required to be satisfied by each type of the heart image were defined and on the basis of number of criteria satisfied by it, each image was evaluated from 0 to 100%. However, for General impression score, no criteria was used and it was evaluated by the experts, on a scale of 0 to 4, on the basis of their experience and understanding.

To automate this manual assessment process of the heart images in TEE, two very well known CNNs (deep learning architectures), AlexNet and VGG16, were implemented. Since the prediction of ”continuous” scores, as close as possible to the manual scores given by the experts, was to be done, two regression models for

regressing over each type of score were built for each network. A classification model, to classify each of the heart image to any one of the ten types of the heart planes, was also constructed for both the networks. These models were trained with original images and the manually evaluated aforementioned scores. For further investigation, these models were experimented with four different processed versions of original images. In total 44 different combinations of models and input images were trained on the GPU (graphical processing unit).

5.2 Conclusions

The objectives set at the beginning of this thesis were met successfully. It can be concluded that CNNs do have the potential to replicate and automate the manual assessment process of imaging skills in TEE which was the primary goal of the project. The regression performed using AlexNet and VGG16 predicted the scores for the heart images, which were quiet close to the manual scores given by the experts. The Root Mean Square Error (RMSE) for predicting criteria percentage score (0 to 100%) was in the range 16.80-19.10 % for AlexNet and 9.64-12.51% for VGG16 **a**nd for predicting general impression score (0 to 4), it was in the range of 0.77-0.87 for AlexNet and 0.40-0.56 for VGG16. After changing the scale for general impression score from 0-4 to 0-100%, it was found that the RMSE for both the scores is less than 20% which is quite low. This can be explained as follows: Since all the 10 planes of the heart either had 4, 5 or 6 criteria associated with them, which were to be satisfied to obtain a perfect view, it implies that approximately 20% corresponds to one criteria being satisfied. As the error obtained is less than 20 % , it implies that there can be an error of maximum one criteria. It was also seen that the estimation of error when performing regression with the criteria percentage score as label was lesser as compared to error when performing regression over general impression score. This is quiet logical as the criteria percentage score is more accurate and standard since it is given on the basis of a number of explicitly defined criteria, required to be satisfied while imaging the heart along the desired 10 planes while the general impression score is given just on the basis of the under-

standing and experience of the experts. On doing regression with criteria percentage and general impression score as labels, VGG16 performs better than AlexNet by 7.1607% and 6.19% respectively. This can be explained as VGG16 is much deeper network and has 8 more convolutional layers compared to AlexNet, which are the primary frameworks in a CNN responsible for detecting features from input images. For each of the heart plane there are certain criteria associated to obtain the perfect view for that plane. For example, Mid-Esophageal 4-Chamber (centered at tricuspid valve) - ME4C (TV) (plane number 1) requires Tricuspid Valve to be in the center (Fig. 5.1) [4]. Mid-Esophageal Aortic Valve Short-Axis - ME AV SAX (plane number 3) requires Aortic Valve of the heart to be in the center of the screen (Fig. 5.2) and so on [4]. Features from such criteria based shapes and templates are better extracted in VGG16 and hence prediction of criteria based score is closer to the manual scores.

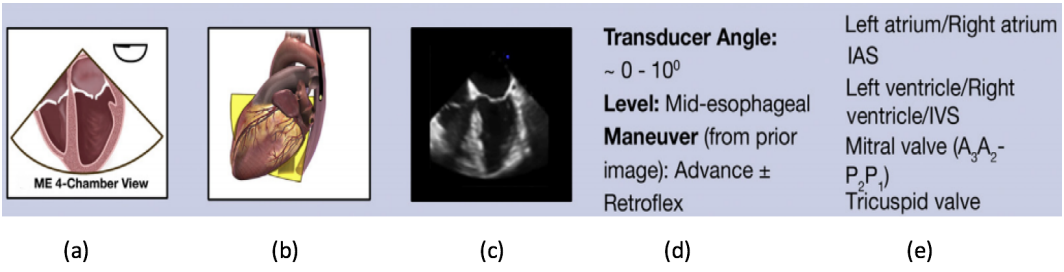


Figure 5.1: (a) Imaging plane; (b) 3D model; (c) 2D TEE image; (d) Acquisition protocol; (e) Structures imaged for Mid-Esophageal 4-Chamber (centered at tricuspid valve) - ME4C (TV) (plane number 1). Image adapted from [4]

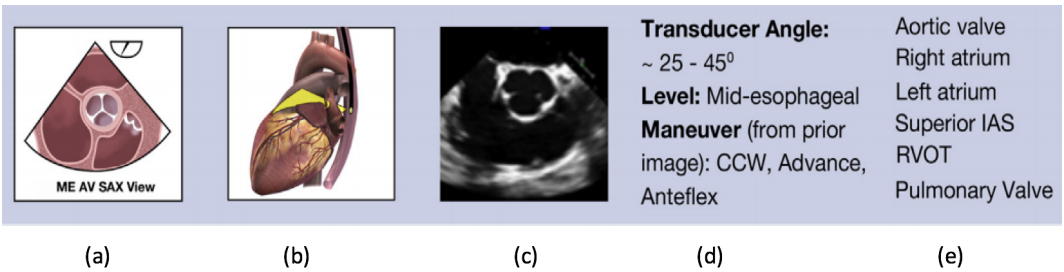


Figure 5.2: (a) Imaging plane; (b) 3D model; (c) 2D TEE image; (d) Acquisition protocol; (e) Structures imaged for Mid-Esophageal Aortic Valve Short-Axis - ME AV SAX (plane number 3). Image adapted from [4]

All types of the input images for performing regression over criteria percent-

age score and general impression score are giving results on the similar lines except the canny images which are giving relatively a slightly higher RMSE values as after applying canny edge detector algorithm most of the features aren't visible clearly and hence feature extraction is not as good as compared to the other input images. There is not much difference when performing classification with both the networks. VGG16 is performing only 0.311% better than AlexNet when classifying heart images into the 10 different planes of the heart from which they were acquired. Hence, in case of classification it is better to implement AlexNet as it is computationally less expensive and takes comparatively less time for training.

Thus, the primary objective to automate the manual assessment process of images obtained by echocardiographers performing TEE has been achieved. Consequently, comments on the imaging skills of an echocardiographer can also be made after precisely predicting the scores for the ultrasound heart images using CNNs for each type of the heart plane. Further, the other objectives of comparing the networks after predicting the scores through AlexNet and VGG16 for original and processed versions of input images, have also been accomplished.

5.3 Future research directions

As the extension of present work, it would be worthwhile to explore the following:

- **Performance Evaluation of both CNNs using a balanced dataset:** The dataset obtained from HeartWorks after the experimentation was unbalanced. For example, their were 3608 out of 16104 images with 100 as criteria percentage score (22% of the dataset), 88 images with 41.667 as score (0.54% of the dataset). Likewise their were unequal number of images for other score values too with the minimum number of images i.e. 44 for 0 as score (0.27%). The performance of a CNN largely depends on the dataset given, if it is trained with the same number images for each of the distinct value of the score, it will predict much better. If the balancing of the dataset was done, i.e. making equal number of images for each of the distinct value of the score which is possible only if for each distinct value, their were 44 images considering the

minimum image number for the 0 score value, then only 9.28% of the dataset i.e. 1496 images out of 16104 would be left for training the CNNs which is very less.

- **Performance Evaluation of CNNs using images obtained from real TEE test:** In this project, the training and testing of the CNNs is done using the heart images obtained from TEE test using simulation based techniques. But, the actual performance of the CNNs will be authenticated when the testing is done using the heart images obtained from the real TEE test (which were not available for this project) and not the ones obtained from simulation based techniques.

- **Combination of kinematics based features with image based features in a TEE:**

  Till now, the improvement of trainees during their training has only been assessed on the basis of how fluently they handled the TEE probe and not on how precisely the recommended views of the heart were taken by them. Their are few kinematics based parameters summarized in section 2.2 with which the handling of the probe was assessed [7]. In this project, an automatic system was built which could assess the heart views taken by the echocardiographers. A possible future approach could be to build an artificially intelligent system which could combine the assessment of kinematic features of the probe with the assessment of its image-based features in a complete TEE assessment system.

Bibliography

[1] Dr. Abdulla M.. Abdulla. Transesophageal echocardiography (tee) test, http://www.heartsite.com/html/tee.html. *Welcome to HeartSite.com*.

[2] Transoesophageal echocardiogram, http://www.learnabouttravelmaps.info/pics/g/gastroscopie-met-roesje-ervaringen.html. *Heartscope Victoria*.

[3] Echocardiogram, http://www.nhs.uk/conditions/echocardiogram/pages/ introduction.aspx. *NHS Choices*.

[4] Jack S Shanewise, Albert T Cheung, Solomon Aronson, William J Stewart, Richard L Weiss, Jonathan B Mark, Robert M Savage, Pamela Sears-Rogan, Joseph P Mathew, Miguel A Quiñones, et al. Ase/sca guidelines for performing a comprehensive intraoperative multiplane transesophageal echocardiography examination: recommendations of the american society of echocardiography council for intraoperative echocardiography and the society of cardiovascular anesthesiologists task force for certification in perioperative transesophageal echocardiography. *Anesthesia & Analgesia*, 89(4):870, 1999.

[5] Heartworks elearn, http://learn.heartworks.me.uk/courses. *Course Information*.

[6] Rebecca T Hahn, Theodore Abraham, Mark S Adams, Charles J Bruce, Kathryn E Glas, Roberto M Lang, Scott T Reeves, Jack S Shanewise, Samuel C Siu, William Stewart, et al. Guidelines for performing a comprehensive transesophageal echocardiographic examination: recommendations from the american society of echocardiography and the society of cardiovas-

cular anesthesiologists. *Journal of the American Society of Echocardiography*, 26(9):921–964, 2013.

[7] Evangelos B Mazomenos, Francisco Vasconcelos, Jeremy Smelt, Henry Prescott, Marjan Jahangiri, Bruce Martin, Andrew Smith, Susan Wright, and Danail Stoyanov. Motion-based technical skills assessment in transoesophageal echocardiography. In *International Conference on Medical Imaging and Virtual Reality*, pages 96–103. Springer, 2016.

[8] Samer Hijazi, Rishi Kumar, and Chris Rowen. Using convolutional neural networks for image recognition. 2015.

[9] Kdnuggets. *KDnuggets Analytics Big Data Data Mining and Data Science*.

[10] Adit Deshpande. A beginners guide to understanding convolutional neural networks. *A Beginners Guide To Understanding Convolutional Neural Networks Adit Deshpande CS Undergrad at UCLA (19)*.

[11] Aarshay Jain, Faizan Shaikh, Sunil Ray, and Shantanu Kumar. Deep learning for computer vision - introduction to convolution neural networks. *Analytics Vidhya*, May 2017.

[12] Adit Deshpande. A beginners guide to understanding convolutional neural networks part 2. *A Beginners Guide To Understanding Convolutional Neural Networks Part 2 Adit Deshpande CS Undergrad at UCLA (19)*.

[13] Sunil Ray, Faizan Shaikh, and Shantanu Kumar. 7 types of regression techniques you should know. *Analytics Vidhya*, May 2017.

[14] Andre Esteva. Skin cancer classification with deep learning. *Dermatologist-level classification of skin cancer with deep neural networks*.

[15] Tizita Nesibu Shewaye. Cardiac MR image segmentation techniques: an overview. *CoRR*, abs/1502.04252, 2015.

[16] Krzysztof J. Geras, Stacey Wolfson, S. Gene Kim, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *CoRR*, abs/1703.07047, 2017.

[17] E.b. mazomenos, f. vasconcelos, j. smelt, m. jahangiri, b. martin, a. smith, s. wright and d. stoyanov. "objectively assessing performance in transoesophageal echocardiography from image comparison and alignment". in 7th joint workshop on new technologies for computer/robot assisted surgery (cras 2017), september 14-15, 2017, montpellier, france (accepted).

[18] Andrej Karpathy. Cs231n convolutional neural networks for visual recognition– module 1, part 2. *CS231n Convolutional Neural Networks for Visual Recognition*, 2015.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[20] Jack Burdick, Oge Marques, Adrià Romero López, Xavier Giró-i Nieto, and Janet Weinthal. Siim 2017 scientific session analytics & deep learning part 3.

[21] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *CoRR*, abs/1702.05747, 2017.

[22] Afonso Menegola, Michel Fornaciali, Ramon Pires, Sandra Eliza Fontes de Avila, and Eduardo Valle. Towards automated melanoma screening: Exploring transfer learning schemes. *CoRR*, abs/1609.01228, 2016.

[23] Miguel A Quiñones, Pamela S Douglas, Elyse Foster, John Gorcsan, Jannet F Lewis, Alan S Pearlman, Jack Rychik, Ernesto E Salcedo, James B Seward, J Geoffrey Stevenson, et al. Acc/aha clinical competence statement on

echocardiography123. *Journal of the American Society of Echocardiography*, 16(4):379–402, 2003.

[24] Who needs transesophageal echocardiography? *National Heart Lung and Blood Institute*, Feb 2016.

[25] Miguel A Quiñones, Pamela S Douglas, Elyse Foster, John Gorcsan, Jannet F Lewis, Alan S Pearlman, Jack Rychik, Ernesto E Salcedo, James B Seward, J Geoffrey Stevenson, et al. Acc/aha clinical competence statement on echocardiography123. *Journal of the American Society of Echocardiography*, 16(4):379–402, 2003.

[26] Werner G Daniel, Raimund Erbel, Wolfgang Kasper, Cees A Visser, Rolf Engberding, George R Sutherland, Eberhard Grube, Peter Hanrath, Bernhard Maisch, and Karl Dennig. Safety of transesophageal echocardiography. a multicenter survey of 10,419 examinations. *Circulation*, 83(3):817–821, 1991.

[27] Jan N Hilberath, Daryl A Oakes, Stanton K Shernan, Bernard E Bulwer, Michael N DAmbra, and Holger K Eltzschig. Safety of transesophageal echocardiography. *Journal of the American Society of Echocardiography*, 23(11):1115–1127, 2010.

[28] S K Mathur and Pooja Singh. Transoesophageal echocardiography related complications. *Indian Journal of Anaesthesia*, Oct 2009.

[29] Cyril Charron, Gwenaël Prat, Vincent Caille, Guillaume Belliard, Montaine Lefèvre, Philippe Aegerter, Jean-Michel Boles, François Jardin, and Antoine Vieillard-Baron. Validation of a skills assessment scoring system for transesophageal echocardiographic monitoring of hemodynamics. *Intensive care medicine*, 33(10):1712–1718, 2007.

[30] Carol E Reiley, Henry C Lin, David D Yuh, and Gregory D Hager. Review of methods for objective surgical skill evaluation. *Surgical endoscopy*, 25(2):356–366, 2011.

[31] Evangelos B Mazomenos, Ping-Lin Chang, Alexander Rolls, David J Hawkes, Colin D Bicknell, Emmanuel Vander Poorten, Celia V Riga, Adrien Desjardins, and Danail Stoyanov. A survey on the current status and future challenges towards objective skills assessment in endovascular surgery. *Journal of Medical Robotics Research*, 1(03):1640010, 2016.

[32] Mark S Adams. Teaching tee for use in the operating room: where are things now... and where are we going? *Journal of the American Society of Echocardiography: official publication of the American Society of Echocardiography*, 25(6):17A, 2012.

[33] Jeremy Smelt, Carlos Corredor, Mark Edsell, Nick Fletcher, Marjan Jahangiri, and Vivek Sharma. Simulation-based learning of transesophageal echocardiography in cardiothoracic surgical trainees: a prospective, randomized study. *The Journal of thoracic and cardiovascular surgery*, 150(1):22–25, 2015.

[34] Supervised and unsupervised machine learning algorithms. *Machine Learning Mastery*, Sep 2016.

[35] Cs231n convolutional neural networks for visual recognition. *CS231n Convolutional Neural Networks for Visual Recognition*.

[36] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[37] Alexey Grigorevich Ivakhnenko. Polynomial theory of complex systems. *IEEE transactions on Systems, Man, and Cybernetics*, 1(4):364–378, 1971.

[38] KS GGCS. Learning representations by back-propagating errors. *Nature*, 323(9), 1986.

[39] Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern recognition*, 15(6):455–469, 1982.

[40] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. pages 267–285, 1982.

[41] Yoshua Bengio Ian Goodfellow and Aaron Courville. Chapter 9 (convolutional networks). *Deep Learning*.

[42] Unsupervised feature learning and deep learning tutorial. *Unsupervised Feature Learning and Deep Learning Tutorial*.

[43] Andrej Karpathy. Cs231n convolutional neural networks for visual recognition (backpropogation). *CS231n Convolutional Neural Networks for Visual Recognition*.

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Intelligent Signal Processing*, pages 306–351. IEEE Press, 1998.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[46] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[50] Andrej Karpathy. Cs231n convolutional neural networks for visual recognition– module 2 (transfer learning). *CS231n Convolutional Neural Networks for Visual Recognition*, 2015.

[51] Greg Chu. How to use transfer learning and fine-tuning in keras and tensorflow to build an image recognition... *Deep Learning Sandbox*, May 2017.

[52] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

[53] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric CNN regression. *CoRR*, abs/1703.07834, 2017.

[54] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. pages 2938–2946, 2015.

[55] Albert Chon, Niranjan Balachandar, and Peter Lu. Deep convolutional neural networks for lung cancer detection.

[56] Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, and Vijay Chandrasekhar. Deep learning for lung cancer detection: Tackling the kaggle data science bowl 2017 challenge. *arXiv preprint arXiv:1705.09435*, 2017.

[57] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. pages 411–418, 2013.

[58] S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong Ki Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[59] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

[60] Adria Romero Lopez, Xavier Giro-i Nieto, Jack Burdick, and Oge Marques. Skin lesion classification from dermoscopic images using deep learning techniques. pages 49–54, 2017.

[61] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[62] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[63] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.